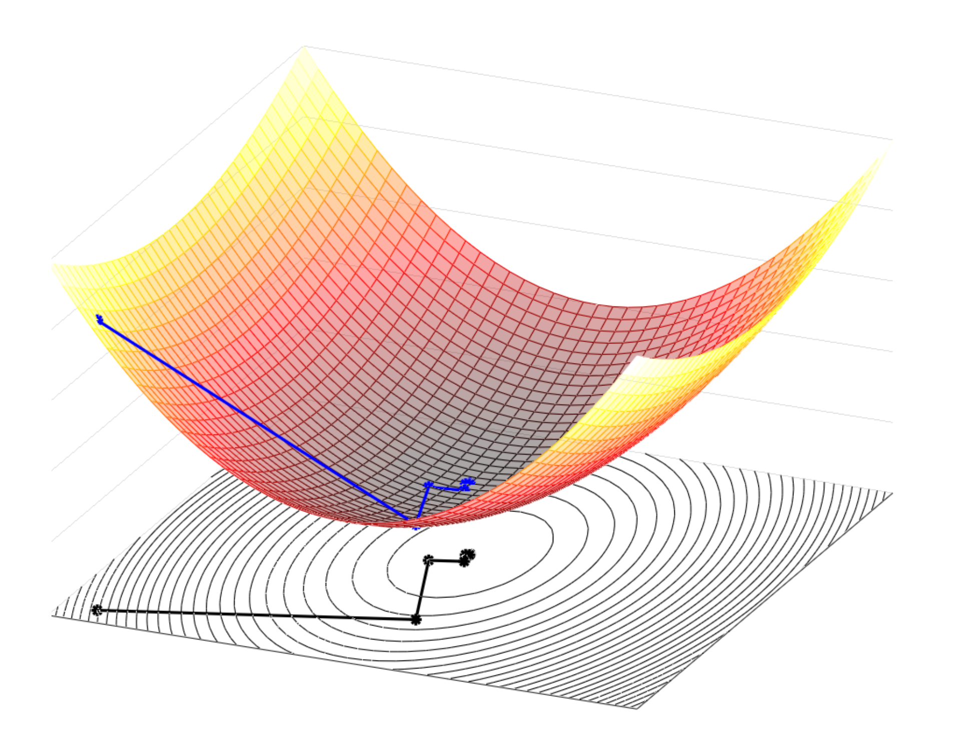
Universidad de El Salvador

Facultad de Ciencias Naturales y Matemática
Escuela de Matemática



Introducción a la optimización numérica matricial y primeros algoritmos

---

Materia
Seminario de Matemática

Ciclo
II-2021

Docente
Dr. Simón Alfredo Peña Aguilar

Estudiante
Br. Leonel Antonio Prudencio Castro PC11012

Fecha
24 de septiembre de 2023

1. Índice

Índice

1. Índice **2**

2. Introducción **3**
2.1. Temas a desarrollar 3
2.2. Prerrequisitos 3

3. Objetivos **4**
3.1. Objetivo general 4
3.2. Objetivos específicos 4

4. Planificación **5**
4.1. Cronograma 5

5. Preliminares **6**
5.1. Definiciones y resultados 6
5.2. El teorema de proyección; primeras consecuencias 7
5.3. Ejercicios 11

6. El problema de optimización **16**
6.1. Generalidades del problema de optimización 16
6.2. Ejemplos de problemas de optimización 22
6.3. Ejercicios 25

7. Métodos de optimización **35**
7.1. Métodos de relajación y gradiente para problemas sin restricciones 35
7.2. Métodos de gradiente conjugado para problemas sin restricciones 46
7.3. Métodos de relación, gradiente y de penalización para problemas con restricciones 54
7.4. Ejercicios 60

8. Conclusiones **72**

9. Bibliografía **73**

2. Introducción

Dada una función $f : \mathbb{R} \longrightarrow \mathbb{R}$, se busca un z tal que:

$$z \in S \subseteq \mathbb{R}, \text{ y } f(z) = \inf_{x \in S} f(x)$$

; este es el planteamiento básico de los problemas de optimización, luego se establecen las condiciones necesarias y suficientes de los métodos numéricos para resolverlos, por ejemplo que la función f sea dos veces diferenciable.

En este material se estudian funciones de tipo $J : V \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}$, que se llamarán *funcionales*, con lo cual el problema es encontrar un $\mathbf{u}$ tal que:

$$\mathbf{u} \in U \subseteq \mathbb{R}^n, \text{ y } J(\mathbf{u}) = \inf_{\mathbf{v} \in U} J(\mathbf{v}) \tag{P}$$

; para resolver el problema (P) se considera el *teorema de la proyección* y sus consecuencias lo que brindará recursos importantes para lo que sigue, después se realiza una caracterización de las condiciones que garanticen la solución, por ejemplo establecer las características del conjunto U, esto se muestra en el apartado llamado: *generalidades de la optimización.*

Luego se observan *ejemplos de problemas de optimización* según las situaciones consideradas, por ejemplo que el funcional tenga dimensión finita, dimensión infinita, que sea cuadrático o elíptico, etc.

A continuación se construyen algoritmos que resuelvan el problema (P), es decir generar alguna sucesión $\{\mathbf{u}_k\}_{k \geq 0}$ de elementos de U tales que $\lim_{k\to\infty} \mathbf{u}_k = \mathbf{u}$, se estudian: *métodos de relajación, gradiente* y *gradiente conjugado.*

Al problema (P) se le pueden añadir otras condiciones adicionales que se llamarán *restricciones* y se estudian: *métodos de relajación , gradiente y de penalización con restricciones.*

Con todo esto se tendrá las nociones básicas del planteamiento teórico y métodos de solución del problema de optimización en el análisis numérico-matricial.

2.1. Temas a desarrollar

- El teorema de la proyección; primeras consecuencias
- Generalidades del problema de optimización
- Ejemplos de problemas de optimización
- Métodos de relajación y gradiente para problemas sin restricciones
- Métodos de gradiente conjugado para problemas sin restricciones
- Métodos de relajación, gradiente y de penalización para problemas con restricciones

2.2. Prerrequisitos

Es necesario tener nociones básicas en las siguientes materias:

- Álgebra lineal
- Análisis funcional

3. Objetivos

3.1. Objetivo general

Estudiar teóricamente las generalidades del problema de optimización para funcionales y sus métodos de solución numérico-matriciales.

3.2. Objetivos específicos

- Conocer las variantes en el planteamiento del problema de optimización.
- Estudiar las condiciones necesarias y suficientes para la solución del problema de optimización para funcionales en diferentes situaciones.
- Estudiar los métodos de solución numérico-matriciales del problema de optimización considerando las variantes del problema planteado.

4. Planificación

4.1. Cronograma

TABLA

5. Preliminares

5.1. Definiciones y resultados

Sea V un espacio vectorial en el campo $\mathbb{R}$. Un *producto escalar* en V es una función $\langle\cdot,\cdot\rangle : V \times V \longrightarrow \mathbb{R}$ bilineal, simétrica y definida positiva, es decir que satisface:

$$\begin{array}{rcl} \langle \mathbf{u},\cdot\rangle : V \longrightarrow \mathbb{R} & \text{es lineal para todo} & \mathbf{u} \in V \\ \langle \cdot,\mathbf{v}\rangle : V \longrightarrow \mathbb{R} & \text{es lineal para todo} & \mathbf{v} \in V \\ \langle \mathbf{u},\mathbf{v}\rangle = \langle \mathbf{v},\mathbf{u}\rangle & \text{para todo} & \mathbf{u},\mathbf{v} \in V \\ \langle \mathbf{v},\mathbf{v}\rangle = 0 \Leftrightarrow \mathbf{v} = \mathbf{0}, & \text{y } \langle \mathbf{v},\mathbf{v}\rangle \geq 0 & \text{para todo } \mathbf{v} \in V \end{array}$$

Se llama *espacio prehilbertiano* a un espacio provisto de un producto escalar. La aplicación $||\cdot||$ definida por:

$$||\mathbf{v}|| = \sqrt{\langle \mathbf{v},\mathbf{v}\rangle} \text{ para todo } \mathbf{v} \in V$$

es una norma en el espacio V, siempre se considera que un espacio prehilbertiano está equipado con esta norma, lo que también lo convierte en un espacio vectorial normalizado. Si está completo para esta norma, es un *espacio de Hilbert*. Al estar completo cualquier espacio vectorial normado de dimensión finita como el espacio $\mathbb{R}^n$ dotado del producto escalar euclidiano es un ejemplo del espacio de Hilbert.

Se considera de pasada la *desigualdad de Schwarz*:

$$|\langle \mathbf{u},\mathbf{v}\rangle| \leq ||\mathbf{u}||\,||\mathbf{v}|| \text{ para todo } \mathbf{u},\mathbf{v} \in V$$

que se utiliza en particular para probar la desigualdad triangular de la norma asociada con el producto escalar. La desigualdad de Cauchy-Schwarz para el producto escalar euclidiano o la desigualdad de Cauchy-Schwarz para las funciones:

$$\left|\int_I \mathbf{u}\mathbf{v}dx\right| \leq \left(\int_I |\mathbf{u}|^2\,dx\right)^{\frac{1}{2}} \left(\int_I |\mathbf{v}|^2\,dx\right)^{\frac{1}{2}}, I \in \mathbb{R}$$

son casos especiales. Notar que la desigualdad de Schwarz implica la continuidad del producto escalar, considerado como la aplicación del producto $V \times V \longrightarrow \mathbb{R}$. Finalmente, notar que esta desigualdad se convierte en una igualdad si, y solo si, los dos vectores que aparecen en ella son linealmente dependientes.

5.2. El teorema de proyección; primeras consecuencias

• ***Teorema 1 (de proyección):*** Sea U un subconjunto cerrado, convexo y no vacío de un espacio de Hilbert V. Dado cualquier elemento $\mathbf{w} \in V$, existe uno y solo un elemento $P\mathbf{w}$ tal que:

$$P\mathbf{w} \in U \text{ y } ||\mathbf{w} - P\mathbf{w}|| = \inf_{\mathbf{v} \in U} ||\mathbf{w} - \mathbf{v}|| \tag{1}$$

Este elemento $P\mathbf{w} \in U$ comprueba

$$\langle P\mathbf{w} - \mathbf{w}, \mathbf{v} - P\mathbf{w} \rangle \geq 0 \text{ para todo } \mathbf{v} \in U \tag{2}$$

y a la inversa, si un elemento $\mathbf{u}$ satisface

$$\mathbf{u} \in U \text{ y } \langle \mathbf{u} - \mathbf{w}, \mathbf{v} - \mathbf{u} \rangle \geq 0 \text{ para todo } \mathbf{v} \in U$$

entonces $\mathbf{u} = P\mathbf{w}$
La función $P : V \longrightarrow U$ así definida es tal que:

$$||P\mathbf{w}_1 - P\mathbf{w}_2|| \leq ||\mathbf{w}_1 - \mathbf{w}_2|| \text{ para todo } \mathbf{w}_1, \mathbf{w}_2 \in V \tag{3}$$

Finalmente, la función $P : V \longrightarrow U \subset V$ es lineal si y solo si el subconjunto U es un subespacio vectorial, en cuyo caso las desigualdades (2) se reemplaza por igualdades:

$$\langle P\mathbf{w} - \mathbf{w}, \mathbf{v} \rangle \text{ para todo } \mathbf{v} \in U \tag{4}$$

∘ *Observaciones:*

(*i*) La función $P : V \longrightarrow U$ se denomina *operador de proyección*, y el elemento $P\mathbf{w}$ se denomina *proyección del elemento* $\mathbf{w}$ (en el conjunto U), siendo clara la interpretación geométrica de la relación en (1) (*figura 1*), es decir: el elemento "proyectado" $P\mathbf{w}$ es de hecho el elemento del conjunto U "más cercano" al punto $\mathbf{w}$. Asimismo, las desigualdades en (2) reflejan la necesidad intuitivamente obvia de que el ángulo formado por los vectores $P\mathbf{w} - \mathbf{w}$ y $\mathbf{v} - P\mathbf{w}$ sea menor o igual a $\pi/2$ para todos los elementos $\mathbf{v} \in U$ (*figura 1*).

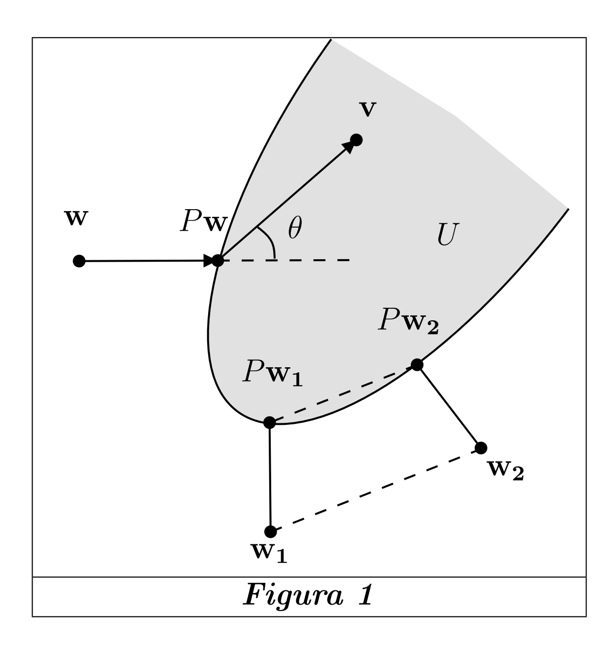


Figura 1

Se observa de pasada que $\mathbf{w} - P\mathbf{w} = \mathbf{0} \Leftrightarrow \mathbf{w} \in U$

(*ii*) La desigualdad (3) conduce en particular a la continuidad del operador de proyección. A veces se retiene diciendo pictóricamente que "la proyección no aumenta las distancias" (*figura 1*).

(*iii*) La condición (4) refleja la ortogonalidad (en el sentido que se definirá más adelante) del vector $P\mathbf{w} - \mathbf{w}$ y de los vectores del conjunto U, cuando este último es un espacio vectorial. La interpretación geométrica aún es evidente (*figura 2*).

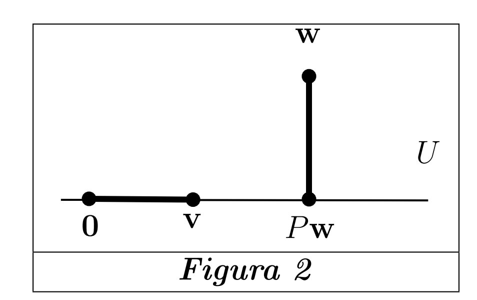


Figura 2

□

Un ejemplo de operador de proyección no lineal en $\mathbb{R}^n$ para $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ es el siguiente:

$$\begin{aligned} V \in \mathbb{R}^n \quad & ; \text{dotado del producto escalar euclideano} \\ U = \mathbb{R}^n_+ \overset{\text{def}}{=} \; & \{\mathbf{u} \in \mathbb{R}^n / u_i \geq 0, 1 \leq i \leq n\} \end{aligned}$$

, el conjunto U a veces se denomina *hiperoctante positivo*. Es casi geométricamente obvio que el operador de proyección correspondiente para $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ y $P\mathbf{w} = ((P\mathbf{w})_1, (P\mathbf{w})_2, \ldots, (P\mathbf{w})_n)$ está definido por:

$$(P\mathbf{w})_i = \text{máx}\{w_i, 0\}, 1 \leq i \leq n$$

como sugiere el examen de todos los "casos" en la dimensión dos (*figura 3*). Para probarlo, basta con verificar la condición necesaria y suficiente del *teorema de proyección*.

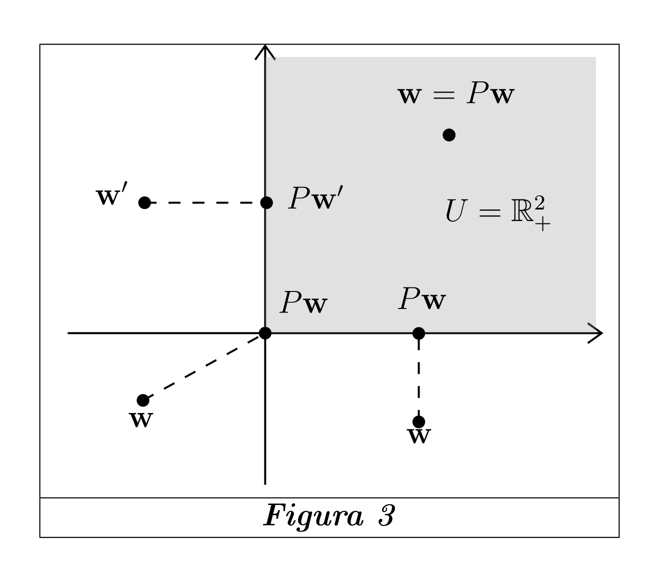


Figura 3

Sin embargo, dado cualquier elemento del conjunto U, la definición previa del elemento $P\mathbf{w}$ implica efectivamente:

$$\langle P\mathbf{w}-\mathbf{w},\mathbf{v}-P\mathbf{w}\rangle=\sum_{i=1}^{n}\left((P\mathbf{w})_i-w_i\right)\left(v_i-(P\mathbf{w})_i\right)=-\sum_{i,w_i<0}w_iv_i\geq 0$$

Es decir, se tiene un conjunto de la forma:

$$U=\prod_{i=1}^{n}[a_i,b_i]=\{\mathbf{v}=(v_1,v_2,\ldots,v_n)\in\mathbb{R}^n/a_i\leq v_i\leq b_i,1\leq i\leq n\}\subset\mathbb{R}^n$$

los casos en los que $a_i=-\infty$ y/o $b_i=\infty$ no están excluidos, no ofrecen ninguna dificultad.

Mediante un razonamiento análogo, operador de proyección correspondiente viene dado por:

$$(P\mathbf{w})_i=\text{mín}\left\{\text{máx}\left\{w_i,a_i\right\},b_i\right\}=\begin{cases}a_i \text{ si} & w_i<a_i\\ w_i \text{ si} & a_i\leq w_i\leq b_i\\ b_i \text{ si} & b_i<w_i\end{cases}$$

Como primera aplicación del teorema de la proyección, se retoma el problema de la solución de un sistema lineal en el sentido de mínimos cuadrados: encuentre $\mathbf{u}\in\mathbb{R}^n$ tal que:

$$||B\mathbf{u}-\mathbf{c}||_m=\inf_{\mathbf{v}\in\mathbb{R}^n}||B\mathbf{v}-\mathbf{c}||_m$$

donde se dan la matriz $B\in\mathcal{A}_{m,n}(\mathbb{R})$ y el vector $\mathbf{c}\in\mathbb{R}^m$, y $||\cdot||_m$ denota la norma euclidiana en $\mathbb{R}^m$. El subespacio vectorial:

$$Im\,(B)=\{B\mathbf{v}\in\mathbb{R}^m/\mathbf{v}\in\mathbb{R}^n\}$$

es cerrado (estamos en dimensión finita), luego el *teorema de proyección* implica la existencia, y la unicidad, de un elemento $\tilde{\mathbf{u}}$ que satisface:

$$\tilde{\mathbf{u}}\in Im\,(B)\ \text{ y }\ ||\tilde{\mathbf{u}}-\mathbf{c}||_m=\inf_{\tilde{\mathbf{v}}\in Im(B)}||\tilde{\mathbf{v}}-\mathbf{c}||_m$$

En consecuencia, el problema planteado siempre tiene al menos una solución, a saber, uno de los elementos $\mathbf{u}\in\mathbb{R}^n$ que verifica:

$$B\mathbf{u}=\tilde{\mathbf{u}}$$

Esta solución es única si y solo si el operador representado por la matriz B es inyectivo (lo cual solo es posible si $m>n$), es decir, si y solo si se define la matriz simétrica positiva $B^{\perp}B$, o nuevamente si y solo si $r(B)=n$.

Con el mismo espíritu, la caracterización (4) del *teorema de proyección*, a saber:

$$\langle\tilde{\mathbf{u}}-\mathbf{c},\tilde{\mathbf{v}}\rangle_m=0 \text{ para todo } \tilde{\mathbf{v}}\in Im\,(B)$$

se escribe, denotando $\langle\cdot,\cdot\rangle_m$ y $\langle\cdot,\cdot\rangle_n$, los productos escalares euclidianos de $\mathbb{R}^m$ y $\mathbb{R}^n$, respectivamente:

$$\langle B\mathbf{u}-\mathbf{c},B\mathbf{v}\rangle_m=\left\langle B^tB\mathbf{u}-B^t\mathbf{c},\mathbf{v}\right\rangle_m=0 \text{ para todo } \mathbf{v}\in\mathbb{R}^n$$

Por tanto, se ha establesido que las ecuaciones normales:

$$B^t B\mathbf{u} = B^t \mathbf{c}$$

siempre tenga al menos una solución.

Dado un elemento $\mathbf{u} \in V$, la desigualdad de Schwarz muestra que a función:

$$\langle \mathbf{u}, \cdot \rangle : \mathbf{v} \in V \longrightarrow \langle \mathbf{u}, \mathbf{v} \rangle \in \mathbb{R}$$

es continuo. Es notable que lo contrario sea cierto si el espacio es completo: cualquier función lineal continua en un espacio de Hilbert puede ser "representada" por un elemento del espacio, como se muestra en el siguiente resultado (cuya prueba se basa en el teorema de proyección):

• ***Teorema 2 (de representación Riesz):*** Sea V un espacio de Hilbert y f cualquier elemento del dual V' de V. Entonces existe un elemento $\tau f \in V$ y solo uno tal que:

$$f(\mathbf{v}) = \langle \tau f, \mathbf{v} \rangle \text{ para todo } \mathbf{v} \in V$$

La aplicación $\tau : V' \longrightarrow V$ así definida es lineal y es una isometría:

$$||\tau f||_V = ||f||_{V'} \text{ para todo } f \in V'$$

La aplicación τ se denomina isometría canónica de Riesz. Una primera aplicación del teorema de representación de Riesz es la extensión de la noción de gradiente: de hecho, si $J : V \longrightarrow \mathbb{R}$ es una función diferenciable en un punto $\mathbf{u}$ de un espacio de Hilbert V, la derivada $J'(\mathbf{v})$ es, por definición, un elemento del dual V'. En consecuencia, existe uno y solo un elemento del espacio V, denotado $\nabla J(\mathbf{u})$, y llamado *gradiente* de la función J en el punto $\mathbf{u}$, tal que:

$$J'(\mathbf{u})\,\mathbf{v} = \langle \nabla J(\mathbf{u}), \mathbf{v} \rangle \text{ para todo } \mathbf{v} \in V$$

Como en la dimensión finita, este vector depende del producto escalar elegido.

De la misma manera, se puede asociar con la segunda derivada $J''(\mathbf{u}) \in \mathcal{L}(V; V')$ un elemento $\nabla^2 J(\mathbf{u})$ del espacio $\mathcal{L}(V)$ tal que:

$$J''(\mathbf{u}) \langle \mathbf{v}, \mathbf{w} \rangle = \left\langle \nabla^2 J(\mathbf{u})\, \mathbf{v}, \mathbf{w} \right\rangle \text{ para todo } \mathbf{v}, \mathbf{w} \in V$$

Dos vectores $\mathbf{u}$ y $\mathbf{v}$ de un espacio prehilbertiano son ortogonales si $\langle \mathbf{u}, \mathbf{v} \rangle = 0$. Si U es cualquier subconjunto de un espacio prehilbertiano V, se denomina: *complemento ortogonal de U* al conjunto:

$$U^\perp \stackrel{\text{def}}{=} \{\mathbf{v} \in V / \langle \mathbf{u}, \mathbf{v} \rangle = 0 \text{ para todo } \mathbf{u} \in U\}$$

Es fácil ver que el conjunto $U^\perp$ sigue siendo un subespacio vectorial cerrado. En el caso de que U también sea un subespacio vectorial cerrado y el espacio esté completo, es posible usando el teorema de la proyección, demostrar el siguiente resultado:

• ***Teorema 3:*** Sea U un subespacio vectorial cerrado de un espacio de Hilbert V. Entonces el espacio V es la suma directa del subespacio y su complemento ortogonal:

$$V = U \oplus U^\perp$$

En otras palabras, cualquier elemento $\mathbf{w} \in V$ se escribe de una y sólo una forma en la forma:

$$\mathbf{w} = \mathbf{u} + \mathbf{u}' \text{ con } \mathbf{u} \in U, \mathbf{u}' \in U^{\perp}$$

Más precisamente, $\mathbf{u} = P\mathbf{w}$ y $\mathbf{u}' = P'\mathbf{w}$, donde P y $P' = I - P$ denotan respectivamente los operadores de proyección en U y $U^{\perp}$.

Dados dos espacios de Hilbert V y W, dotados de productos escalares $\langle \cdot, \cdot \rangle_V$ y $\langle \cdot, \cdot \rangle_W$, el teorema de representación de Riesz permite asociar cualquier operador $A \in \mathcal{L}(V; W)$ con el operador transpuesto $A^{\perp} \in \mathcal{L}(W; V)$ definido por:

$$\langle A\mathbf{v}, \mathbf{w} \rangle_W = \left\langle \mathbf{v}, A^{\perp}\mathbf{w} \right\rangle_V \text{ para todo } \mathbf{v} \in V, \mathbf{w} \in W$$

Naturalmente, se tiene la definición habitual de una matriz transpuesta cuando los espacios V y W son de dimensión finita y se dotan del producto escalar euclidiano. De la definición anterior y el teorema anterior se deducen las relaciones:

$$V = Ker\,(A) \oplus \overline{Im\,(A^{\perp})}, W = Ker\,\left(A^{\perp}\right) \oplus \overline{Im\,(A)}$$

; donde se usan las notaciones habituales

$$Ker\,(A) = \{\mathbf{v} \in V / A\mathbf{v} = \mathbf{0}\}\,, Im\,(A) = \{A\mathbf{v} \in W / \mathbf{v} \in V\}$$

para el núcleo y la imagen, respectivamente, del operador lineal A.

Se usan estas relaciones en el caso particular donde los dos espacios V y W son de dimensión finita, en cuyo caso los subespacios $Im(A)$ e $Im(A^{\perp})$ están siempre cerrados. Estas relaciones llevan a veces el nombre de alternativa de *Fredholm en dimensión finita*, debido a las consecuencias que se deducen para la resolución de un sistema lineal con matriz no necesariamente cuadrada, a saber:

Sean V y W dos espacios de dimensión finita, A un operador lineal de V en W, y $\mathbf{b}$ un vector de W. Entonces ocurre una, y sólo una, de las siguientes dos posibilidades:

- El sistema lineal $A\mathbf{v} = \mathbf{b}$ tiene al menos una solución
- El sistema lineal $A\mathbf{v} = \mathbf{b}$ no tiene solución y existe al menos un vector $\mathbf{w} \in W$ tal que $A^{\perp}\mathbf{w} = 0$ y $\langle \mathbf{w}, \mathbf{b} \rangle \neq 0$ (por ejemplo, la proyección del vector $\mathbf{b}$ en el núcleo de la aplicación transpuesta $A^{\perp}$).

5.3. Ejercicios

Ejercicio 5.1: Sea $\mathbf{v}$ un vector real que verifique: $\mathbf{v}^t\mathbf{v} = 1$. Demuestre que la matriz $(I - \mathbf{v}\mathbf{v}^t)$ representa un operador de proyección. ¿Qué propiedad geométrica resulta para la matriz de Householder $H(\mathbf{v}) = I - 2\mathbf{v}\mathbf{v}^t$?

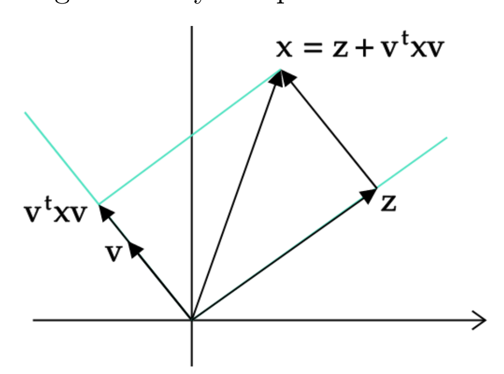
◦ *Solución*

Sea $A = I - \mathbf{v}\mathbf{v}^t$, hay que ver que se cumpla: $A^2 = A$, sea $\mathbf{u} \in \mathbb{R}^n$

$$\begin{aligned}
A^2\mathbf{u} &= AA\mathbf{u} = A\left(I - \mathbf{v}\mathbf{v}^t\right)\mathbf{u} = A\left(I\mathbf{u} - \mathbf{v}\mathbf{v}^t\mathbf{u}\right) = A\mathbf{u} - A\left(\mathbf{v}\mathbf{v}^t\mathbf{u}\right) \\
&= \left(I - \mathbf{v}\mathbf{v}^t\right)\mathbf{u} - \left(I - \mathbf{v}\mathbf{v}^t\right)\mathbf{v}\mathbf{v}^t\mathbf{u} \\
&= \mathbf{u} - \mathbf{v}\mathbf{v}^t\mathbf{u} - \mathbf{v}\mathbf{v}^t\mathbf{u} + \mathbf{v}\mathbf{v}^t\mathbf{v}\mathbf{v}^t\mathbf{u} \\
&= \mathbf{u} - \mathbf{v}\mathbf{v}^t\mathbf{u} - \mathbf{v}\mathbf{v}^t\mathbf{u} + \mathbf{v}\mathbf{v}^t\mathbf{u} \\
&= \left(I - \mathbf{v}\mathbf{v}^t\right)\mathbf{u} \\
&= A\mathbf{u}
\end{aligned}$$

$\Rightarrow A^2 = A$, por lo que es una proyección

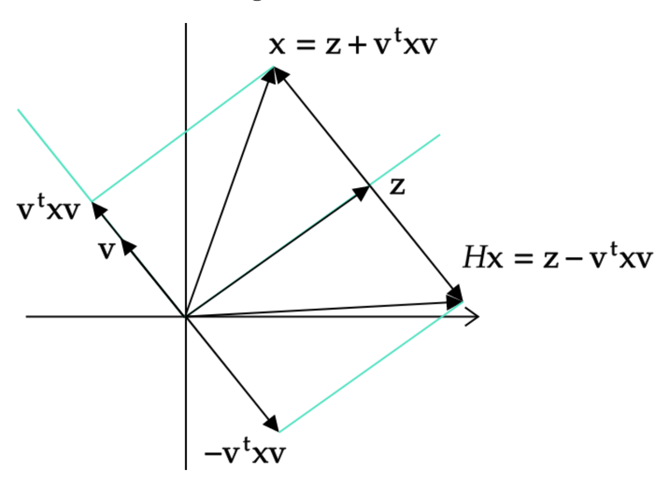
Para la segunda parte del problema: Sean $\mathbf{x}$ y $\mathbf{v} \in \mathbb{R}^n$ tal que $\mathbf{v}^t\mathbf{v} = 1$, se considera la proyección de $\mathbf{x}$ sobre $\mathbf{v}$: $< \mathbf{v}, \mathbf{x} > \mathbf{v} = \mathbf{v}^t\mathbf{x}\mathbf{v}$ ($< \cdot, \cdot >$ denota producto punto) y luego se se toma un vector $\mathbf{z}$ que sea ortogonal a $\mathbf{v}$ y tal que: $\mathbf{x} = \mathbf{z} + \mathbf{v}^t\mathbf{x}\mathbf{v}$



Para $\mathbf{v}^t\mathbf{v} = 1$ y $H = I - 2\mathbf{v}\mathbf{v}^t$, se tiene:

$$\begin{aligned} H\mathbf{x} &= \left(I - 2\mathbf{v}\mathbf{v}^t\right)\mathbf{x} = \left(I - 2\mathbf{v}\mathbf{v}^t\right)\left(\mathbf{z} + \mathbf{v}^t\mathbf{x}\mathbf{v}\right) = \mathbf{z} + \mathbf{v}^t\mathbf{x}\mathbf{v} - 2\mathbf{v}\underbrace{\mathbf{v}^t\mathbf{z}}_{0} - 2\mathbf{v}\mathbf{v}^t\mathbf{v}^t\mathbf{x}\mathbf{v} \\ &= \mathbf{z} + \mathbf{v}^t\mathbf{x}\mathbf{v} - 2\underbrace{\left(\mathbf{v}^t\mathbf{v}\right)^t}_{1}\left(\mathbf{v}^t\mathbf{x}\right)\mathbf{v} = \mathbf{z} - \mathbf{v}^t\mathbf{x}\mathbf{v} \end{aligned}$$

Por lo que $H\mathbf{x}$ es el reflejo de $\mathbf{x}$ sobre el ortogonal de $\mathbf{v}$:



Ejemplo: Para $\mathbf{v} = \begin{bmatrix} 1 & 0 \end{bmatrix}^t$, notar que $\mathbf{v}^t\mathbf{v} = 1$ y sea $\mathbf{x} = \begin{bmatrix} 1 & 2 \end{bmatrix}^t$

$$H = I - 2\begin{bmatrix} 1 \\ 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 2\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Rightarrow H\mathbf{x} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

⊠

Ejercicio 5.2: Sea V un espacio prehilbertiano en el campo $\mathbb{R}$. Demuestre que la norma asociada con el producto escalar satisface la ley del paralelogramo:

$$\|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2 = 2\left(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2\right), \text{ para todo } \mathbf{u}, \mathbf{v} \in V$$

, y que recíprocamente si una norma $||\cdot||$ en un espacio vectorial V satisface la ley del paralelogramo, entonces la función $(\cdot,\cdot) : V \longrightarrow \mathbb{R}$, definida por:

$$2\left(\mathbf{u},\mathbf{v}\right) = \|\mathbf{u}+\mathbf{v}\|^2 - \|\mathbf{u}\|^2 - \|\mathbf{v}\|^2$$

es un producto escalar en V.

$\circ$*Solución:*

$$\begin{aligned}
\|\mathbf{u}+\mathbf{v}\|^2 &= <\mathbf{u}+\mathbf{v},\mathbf{u}+\mathbf{v}>=<\mathbf{u},\mathbf{u}>+<\mathbf{u},\mathbf{v}>+<\mathbf{v},\mathbf{u}>+<\mathbf{v},\mathbf{v}> \\
&= \|\mathbf{u}\|^2+\|\mathbf{v}\|^2+<\mathbf{u},\mathbf{v}>+<\mathbf{v},\mathbf{u}>
\end{aligned}$$

$$\begin{aligned}
\|\mathbf{u}-\mathbf{v}\|^2 &= <\mathbf{u}-\mathbf{v},\mathbf{u}-\mathbf{v}>=<\mathbf{u},\mathbf{u}>-<\mathbf{u},\mathbf{v}>-<\mathbf{v},\mathbf{u}>+<\mathbf{v},\mathbf{v}> \\
&= \|\mathbf{u}\|^2+\|\mathbf{v}\|^2-<\mathbf{u},\mathbf{v}>-<\mathbf{v},\mathbf{u}>
\end{aligned}$$

$\Rightarrow \|\mathbf{u}+\mathbf{v}\|^2+\|\mathbf{u}-\mathbf{v}\|^2 = 2\left(\|\mathbf{u}\|^2+\|\mathbf{v}\|^2\right)$, notar que el espacio V está sobre el campo $\mathbb{R}$.

Hay que ver que $(\mathbf{u},\mathbf{v}) = \frac{1}{2}\left(\|\mathbf{u}+\mathbf{v}\|^2-\|\mathbf{u}\|^2-\|\mathbf{v}\|^2\right)$ es un producto escalar en V

Se sabe que $||\cdot||$ cumple la ley del paralelogramo, es decir: $\|\mathbf{u}+\mathbf{v}\|^2+\|\mathbf{u}-\mathbf{v}\|^2 = 2\left(\|\mathbf{u}\|^2+\|\mathbf{v}\|^2\right)$

$$\begin{aligned}
\Rightarrow \frac{\|\mathbf{u}+\mathbf{v}\|^2+\|\mathbf{u}-\mathbf{v}\|^2}{2} &= \|\mathbf{u}\|^2+\|\mathbf{v}\|^2 \\
\Rightarrow \frac{\|\mathbf{u}+\mathbf{v}\|^2+\|\mathbf{u}-\mathbf{v}\|^2}{4} &= \frac{\|\mathbf{u}\|^2+\|\mathbf{v}\|^2}{2}
\end{aligned}$$

Entonces:

$$\begin{aligned}
(\mathbf{u},\mathbf{v}) &= \frac{\|\mathbf{u}+\mathbf{v}\|^2-\|\mathbf{u}\|^2-\|\mathbf{v}\|^2}{2} = \frac{\|\mathbf{u}+\mathbf{v}\|^2}{2} - \frac{\left(\|\mathbf{u}\|^2+\|\mathbf{v}\|^2\right)}{2} \\
&= \frac{\|\mathbf{u}+\mathbf{v}\|^2}{2} - \frac{\|\mathbf{u}+\mathbf{v}\|^2}{4} - \frac{\|\mathbf{u}-\mathbf{v}\|^2}{4} \\
&= \frac{1}{4}\left(\|\mathbf{u}+\mathbf{v}\|^2-\|\mathbf{u}-\mathbf{v}\|^2\right)
\end{aligned}$$

A continuación se procede a verificar las propiedades del producto escalar.

- $(\mathbf{u},\mathbf{v}) = (\mathbf{v},\mathbf{u})$

$$\begin{aligned}
(\mathbf{u},\mathbf{v}) &= \frac{\|\mathbf{u}+\mathbf{v}\|^2-\|\mathbf{u}\|^2-\|\mathbf{v}\|^2}{2} \\
&= \frac{\|\mathbf{v}+\mathbf{u}\|^2-\|\mathbf{v}\|^2-\|\mathbf{u}\|^2}{2} = (\mathbf{v},\mathbf{u})
\end{aligned}$$

- $(\mathbf{u}+\mathbf{v},\mathbf{w}) = (\mathbf{u},\mathbf{w})+(\mathbf{v},\mathbf{w})$

Por ley del paralelogramo:

$$\begin{aligned} 2\left\|\mathbf{u}+\mathbf{w}\right\|^2+2\left\|\mathbf{v}\right\|^2 &= \left\|\mathbf{u}+\mathbf{v}+\mathbf{w}\right\|^2+\left\|\mathbf{u}-\mathbf{v}+\mathbf{w}\right\|^2 \\ \Rightarrow \left\|\mathbf{u}+\mathbf{v}+\mathbf{w}\right\|^2 &= 2\left\|\mathbf{u}+\mathbf{w}\right\|^2+2\left\|\mathbf{v}\right\|^2-\left\|\mathbf{u}-\mathbf{v}+\mathbf{w}\right\|^2 \\ &= 2\left\|\mathbf{v}+\mathbf{w}\right\|^2+2\left\|\mathbf{u}\right\|^2-\left\|\mathbf{v}-\mathbf{u}+\mathbf{w}\right\|^2 \\ & \quad ;\ \text{intercambiando } \mathbf{u} \text{ con } \mathbf{v} \end{aligned}$$

$$\Rightarrow \left\|\mathbf{u}+\mathbf{v}+\mathbf{w}\right\|^2=\left\|\mathbf{u}\right\|^2+\left\|\mathbf{v}\right\|^2+\left\|\mathbf{u}+\mathbf{w}\right\|^2+\left\|\mathbf{v}+\mathbf{w}\right\|^2-\frac{\left\|\mathbf{u}-\mathbf{v}+\mathbf{w}\right\|^2}{2}-\frac{\left\|\mathbf{v}-\mathbf{u}+\mathbf{w}\right\|^2}{2}$$

$$\Rightarrow \left\|\mathbf{u}+\mathbf{v}-\mathbf{w}\right\|^2=\left\|\mathbf{u}\right\|^2+\left\|\mathbf{v}\right\|^2+\left\|\mathbf{u}-\mathbf{w}\right\|^2+\left\|\mathbf{v}-\mathbf{w}\right\|^2-\frac{\left\|\mathbf{u}-\mathbf{v}-\mathbf{w}\right\|^2}{2}-\frac{\left\|\mathbf{v}-\mathbf{u}-\mathbf{w}\right\|^2}{2}$$

; considerando $-\mathbf{w}$.

Luego:

$$\begin{aligned} (\mathbf{u}+\mathbf{v},\mathbf{w}) &= \frac{1}{4}\left(\left\|\mathbf{u}+\mathbf{v}+\mathbf{w}\right\|^2-\left\|\mathbf{u}+\mathbf{v}-\mathbf{w}\right\|^2\right) \\ &= \frac{1}{4}\Bigg(\left\|\mathbf{u}\right\|^2+\left\|\mathbf{v}\right\|^2+\left\|\mathbf{u}+\mathbf{w}\right\|^2+\left\|\mathbf{v}+\mathbf{w}\right\|^2-\frac{\left\|\mathbf{u}-\mathbf{v}+\mathbf{w}\right\|^2}{2}-\frac{\left\|\mathbf{v}-\mathbf{u}+\mathbf{w}\right\|^2}{2} \\ &\quad -\left\|\mathbf{u}\right\|^2-\left\|\mathbf{v}\right\|^2-\left\|\mathbf{u}-\mathbf{w}\right\|^2-\left\|\mathbf{v}-\mathbf{w}\right\|^2+\frac{\left\|\mathbf{u}-\mathbf{v}-\mathbf{w}\right\|^2}{2}+\frac{\left\|\mathbf{v}-\mathbf{u}-\mathbf{w}\right\|^2}{2}\Bigg) \\ &= \frac{1}{4}\Big(\left\|\mathbf{u}+\mathbf{w}\right\|^2+\left\|\mathbf{v}+\mathbf{w}\right\|^2-\left\|\mathbf{u}-\mathbf{w}\right\|^2-\left\|\mathbf{v}-\mathbf{w}\right\|^2 \\ &\quad -\frac{\left\|\mathbf{u}-\mathbf{v}+\mathbf{w}\right\|^2}{2}+\frac{\left\|\mathbf{u}-\mathbf{v}+\mathbf{w}\right\|^2}{2}-\frac{\left\|\mathbf{v}-\mathbf{u}+\mathbf{w}\right\|^2}{2}+\frac{\left\|\mathbf{v}-\mathbf{u}+\mathbf{w}\right\|^2}{2}\Big) \\ &= \frac{1}{4}\left(\left\|\mathbf{u}+\mathbf{w}\right\|^2-\left\|\mathbf{u}-\mathbf{w}\right\|^2\right)+\frac{1}{4}\left(\left\|\mathbf{v}+\mathbf{w}\right\|^2-\left\|\mathbf{v}-\mathbf{w}\right\|^2\right) \\ &= (\mathbf{u},\mathbf{w})+(\mathbf{v},\mathbf{w}) \end{aligned}$$

- $(\lambda\mathbf{u},\mathbf{v})=\lambda(\mathbf{u},\mathbf{v})$

  Para $\lambda=-1$:

$$\begin{aligned} (-\mathbf{u},\mathbf{v}) &= \frac{1}{4}\left(\left\|-\mathbf{u}+\mathbf{v}\right\|^2-\left\|-\mathbf{u}-\mathbf{v}\right\|^2\right) \\ &= \frac{1}{4}\left(\left\|\mathbf{u}-\mathbf{v}\right\|^2-\left\|\mathbf{u}+\mathbf{v}\right\|^2\right)=-(\mathbf{u},\mathbf{v}) \end{aligned}$$

  Ahora se considera una función $f:V\times V\longrightarrow\mathbb{R}$ definida de esta manera: $f(\mathbf{u},\mathbf{v})=(\mathbf{u},\mathbf{v})=\frac{1}{4}\left(\left\|\mathbf{u}+\mathbf{v}\right\|^2-\left\|\mathbf{u}-\mathbf{v}\right\|^2\right)$, notar que es continua porque la función norma $||\cdot||$

es continua, luego sean $\lambda = \frac{p}{q} \in \mathbb{Q}$ ($q \neq 0$ y $p, q > 0$) y $\mathbf{u}' = \frac{\mathbf{u}}{q}$, entonces:

$$\begin{aligned}
q(\lambda \mathbf{u}, \mathbf{v}) &= q(p\mathbf{u}', \mathbf{v}) \\
&= q\left(\underbrace{\mathbf{u}' + \mathbf{u}' + \cdots + \mathbf{u}'}_{p \text{ veces}}, \mathbf{v}\right) \\
&= q\left[\underbrace{(\mathbf{u}', \mathbf{v}) + (\mathbf{u}', \mathbf{v}) + \cdots + (\mathbf{u}', \mathbf{v})}_{p \text{ veces}}\right] \\
&= qp(\mathbf{u}', \mathbf{v}) \\
&= pq(\mathbf{u}', \mathbf{v}) \\
&= p\left[\underbrace{(\mathbf{u}', \mathbf{v}) + (\mathbf{u}', \mathbf{v}) + \cdots + (\mathbf{u}', \mathbf{v})}_{q \text{ veces}}\right] \\
&= p\left(\underbrace{\mathbf{u}' + \mathbf{u}' + \cdots + \mathbf{u}'}_{q \text{ veces}}, \mathbf{v}\right) \\
&= p(q\mathbf{u}', \mathbf{v})
\end{aligned}$$

Dividiendo por q se tiene: $(\lambda \mathbf{u}, \mathbf{v}) = \lambda(\mathbf{u}, \mathbf{v})$, para todo $\lambda \in \mathbb{Q}$

Por la continuidad de la función f se tiene: $(\lambda \mathbf{u}, \mathbf{v}) = \lambda(\mathbf{u}, \mathbf{v})$, para todo $\lambda \in \mathbb{R}$

- $(\mathbf{u}, \mathbf{u}) \geq 0$ y $(\mathbf{u}, \mathbf{u}) = 0 \Leftrightarrow \mathbf{u} = \mathbf{0}$

$$\begin{aligned}
(\mathbf{u}, \mathbf{u}) &= \tfrac{1}{4}\left(\|\mathbf{u} + \mathbf{u}\|^2 - \|\mathbf{u} - \mathbf{u}\|^2\right) = \frac{1}{4}\left(\|2\mathbf{u}\|^2\right) = \|\mathbf{u}\|^2 \geq 0 \\
&\Rightarrow (\mathbf{u}, \mathbf{u}) = 0 \Leftrightarrow \mathbf{u} = \mathbf{0}
\end{aligned}$$

Con todo lo anterior, la función $2(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} + \mathbf{v}\|^2 - \|\mathbf{u}\|^2 - \|\mathbf{v}\|^2$ es un producto escalar sobre V.

$\boxtimes$

6. El problema de optimización

6.1. Generalidades del problema de optimización

Un problema de optimización toma la siguiente forma: dado un conjunto U no vacío de un espacio vectorial V y una función $J : V \longrightarrow \mathbb{R}$, se trata de encontrar un mínimo de la función J con respecto al conjunto U, es decir, un elemento $\mathbf{u}$ que verifica:

$$\mathbf{u} \in U \subseteq V \text{ y } J(\mathbf{u}) = \inf_{\mathbf{v}\in U} J(\mathbf{v}) \tag{P}$$

◦ *Observación:*

Para la definición del problema (P), por lo tanto, es suficiente conocer la función J sobre el conjunto U, pero en la práctica, generalmente se conoce sobre todo el espacio V.

□

Ahora se especifican algunos puntos sobre la terminología, específicamente la naturaleza de la función J, que usualmente se llama *funcional en optimización*, y del conjunto U.

Se distinguen los problemas sin restricciones cuando $U = V$ y los problemas con restricciones en el caso contrario. Entre los problemas con las restricciones, un caso muy importante en las aplicaciones es el de los conjuntos U de la forma:

$$U = \{\mathbf{v} \in V | \varphi_i(\mathbf{v}) \leq 0, 1 \leq i \leq m', \varphi_i(\mathbf{v}) = 0, m'+1 \leq i \leq m\}$$

las funciones dadas $\varphi_i : V \longrightarrow \mathbb{R}, 1 \leq i \leq m$, se llaman las restricciones del problema. Si $m' = m$, o si $m = 0$, a menudo se dice por abuso del lenguaje que es un problema con *"restricciones-desigualdades"*, o con *"restricciones-igualdades"*, respectivamente.

En ausencia de supuestos adicionales sobre las funciones φ_i y J, en particular con respecto a la convexidad y, con mayor motivo, la linealidad, el problema asociado (P) se denomina *problema de programación no lineal.*

Dado que siempre se puede reemplazar una *"restricción-igualdad"* $\varphi_i = 0$ por las dos *"restricciones-desigualdades"*: $\varphi_i(\mathbf{v}) \leq 0$ y $-\varphi_i(\mathbf{v}) \leq 0$ se limita temporalmente a considerar los únicos problemas con *"restricciones-desigualdades"*, correspondientes, por tanto, a los conjuntos U de la forma:

$$U = \{\mathbf{v} \in V | \varphi_i(\mathbf{v}) \leq 0, 1 \leq i \leq m\}$$

Si las funciones j y φ_i son convexas, se dice que es un *problema de programación convexa*, se nota que el conjunto U es entonces convexo; en efecto:

$$\left.\begin{array}{c}\varphi_i(\mathbf{u}) \leq 0; \varphi_i(\mathbf{v}) \leq 0; \\ \theta \in [0,1]\end{array}\right\} \Rightarrow \varphi_i(\theta\mathbf{u} + (1-\theta)\mathbf{v}) \leq \theta\varphi_i(\mathbf{u}) + (1-\theta)\varphi_i(\mathbf{v}) \leq 0$$

y una intersección de conjuntos convexos es convexa.

Dos casos especiales muy importantes de programación convexa son los de *programación cuadrática* y *programación lineal*; en un problema de programación cuadrática, la función J es un funcional cuadrático en $V = \mathbb{R}^n$:

$$J : \mathbf{v} \in \mathbb{R}^n \longrightarrow J(\mathbf{v}) = \frac{1}{2}\langle A\mathbf{v}, \mathbf{v}\rangle - \langle \mathbf{b}, \mathbf{v}\rangle, A = A^t \in \mathcal{A}_n(\mathbb{R}), \mathbf{b} \in \mathbb{R}^n$$

se supone que la matriz A es definida positiva (lo que implica la convexidad estricta de la función J), y las restricciones φ_i son afines (por lo tanto convexas):

$$U = \left\{ \mathbf{v} \in \mathbb{R}^n | \sum_{j=1}^{n} c_{ij} v_j \leq d_i, 1 \leq i \leq m \right\}$$

En un problema de programación lineal, la función J es una funcional lineal sobre $V = \mathbb{R}^n$:

$$J(\mathbf{v}) = \sum_{i=1}^{n} a_i v_i$$

y el conjunto U todavía tiene la forma:

$$U = \left\{ \mathbf{v} \in \mathbb{R}^n | \sum_{j=1}^{n} c_{ij} v_j \leq d_i, 1 \leq i \leq m \right\}$$

◦ *Observación:*

Si la matriz simétrica que interviene en la definición de un funcional cuadrático es solo positiva, este último sigue siendo convexo; por lo tanto, sería concebible volver a llamar al problema de optimización correspondiente como un problema de programación cuadrática. Sin embargo, al hacerlo, la programación lineal parecería ser un caso especial de programación cuadrática, que es extremadamente inexacta en muchos aspectos, hasta el punto de que se considere a parte el caso de la programación lineal.

□

Se examinan ahora las cuestiones de existencia y unicidad de la solución al problema (P). Ya sea que se el caso de dimensión finita o no, la unicidad de una posible solución generalmente se establece independientemente de la existencia, la mayoría de las veces a partir de la convexidad del conjunto U y la convexidad estricta del funcional.

En cuanto a la existencia, se comienza por el caso de la dimensión finita. Si U es un conjunto acotado y cerrado de $V = \mathbb{R}^n$ y si la función $J : \mathbb{R}^n \longrightarrow \mathbb{R}$ es continua, está claro que el problema (P) tiene al menos una solución. Para extender inicialmente este resultado al caso de conjuntos U no acotados (en particular cuando $U = V = \mathbb{R}^n$), se introduce la siguiente noción: una función J con valores reales definidos en un espacio vectorial normado V se dice *coercitiva* si:

$$\lim_{\|\mathbf{v}\|_V \to \infty} J(\mathbf{v}) = +\infty$$

• ***Teorema 4:*** Sea U una conjunto cerrado no vacío de $\mathbb{R}^n$, y $J : \mathbb{R}^n \longrightarrow \mathbb{R}$ una función continua, coercitiva si el conjunto U es no acotado. Entonces existe al menos un elemento $\mathbf{u}$ tal que:

$$\mathbf{u} \in U \text{ y } J(\mathbf{u}) = \inf_{\mathbf{v} \in U} J(\mathbf{v}) \qquad \text{(P)}$$

◦ *Demostración:*

Sea $\mathbf{u}_0$ cualquier punto del conjunto U. La coercitividad del funcional J implica la existencia de un número r tal que:

$$\|\mathbf{v}\| > r \Rightarrow J(\mathbf{u}_0) < J(\mathbf{v})$$

En estas condiciones, el conjunto de soluciones del problema (P) coincide con el de las soluciones del problema (P_0) correspondiente al conjunto:

$$U_0 = U \cap \{\mathbf{v} \in \mathbb{R}^n | \, \|\mathbf{v}\| \leq r\}$$

Por lo tanto, se ha reducido al caso de un subconjunto no vacío ($\mathbf{u}_0 \in U_0$), cerrado y acotado.

■

○ *Observaciones:*

(*i*) El *teorema 4* proporciona una prueba del teorema de la proyección (*teorema 1*) cuando el espacio V es de dimensión finita; basta con introducir la función (con las notaciones del teorema *teorema 1*) $J(\mathbf{v}) = \|\mathbf{w} - \mathbf{v}\|$ que es coercitivo ya que $J(\mathbf{v}) \geq \|\mathbf{v}\| - \|\mathbf{w}\|$. Pero este punto de vista hace que la compacidad juegue un papel artificial: la demostración del teorema de la proyección se basa por un lado en la completitud del espacio y por otro lado en la "geometría"del espacio, ligada a la existencia de un producto escalar. Por otro lado, la ventaja de la presente demostración es que se aplica a cualquier estándar.

(*ii*) Notar que, cuando el conjunto U no está acotado y el funcional es lineal, el resultado anterior no se aplica generalmente.

□

Es la compacidad la que interviene de manera esencial en la demostración del *teorema 4*. Se puede convencer de lo contrario si se considera una sucesión minimizadora $(\mathbf{u}_k)_{k\geq 0}$, es decir una sucesión de puntos que verifican:

$$\mathbf{u}_k \in, \forall k \geq 0, \lim_{k\to\infty} J(\mathbf{u}_k) = \inf_{\mathbf{v}\in U} J(\mathbf{v})$$

Esta sucesión necesariamente acotada, dado que el funcional J es coercitivo, se puede extraer una subsucesión $(\mathbf{u}_{k'})$ que converge hacia un elemento $\mathbf{u} \in U$ (el conjunto U es cerrado).

La función J es continua,

$$J(\mathbf{u}) = \lim_{k'\to\infty} J(\mathbf{u}_{k'}) = \inf_{\mathbf{v}\in U} J(\mathbf{v})$$

que proporciona una nueva prueba de la existencia de una solución del problema (P).

Es además este tipo de razonamiento el que permite extender el resultado al caso de dimensión infinita, sin embargo con supuestos de convexidad adicionales y esenciales, tanto para el funcional J como para el conjunto U. La prueba basada en la compacidad "débil" de las partes convexas cerradas y acotadas de los espacios de Hilbert (partes (ii) y (iii) de la siguiente demostración), se comienza con la siguiente definición: Se dice que una sucesión $(\mathbf{u}_k)_{k\geq 0}$ de elementos de un espacio prehilbertiano V converge débilmente si existe un elemento $\mathbf{u} \in V$ tal que:

$$\lim_{k\to\infty} \langle \mathbf{v}, \mathbf{u}_k \rangle = \langle \mathbf{v}, \mathbf{u} \rangle, \forall \mathbf{v} \in V$$

Se observará que, si cualquier sucesión que converge dentro del significado de la norma, ésta converge débilmente, lo contrario no siempre es cierto.

• ***Teorema 5:*** Sea U un conjunto cerrado, convexo y no vacío de un espacio de Hilbert separable V, y $J : V \longrightarrow \mathbb{R}$ un funcional convexo, derivable y coercitivo si el conjunto U no está acotado. Entonces existe al menos un elemento $\mathbf{u}$ tal que:

$$\mathbf{u} \in U \text{ y } J(\mathbf{u}) = \inf_{\mathbf{v}\in U} J(\mathbf{v}) \tag{P}$$

∘ *Demostración:*

(*i*) Como en el caso de la dimensión finita (*teorema 4*), la coercitividad del funcional permite volver al caso único de un conjunto acotado U (y nuevamente convexo ya que una bola es convexa; ver la demostración del teorema antes mencionado).

(*ii*) Se considera una secesión minimizadora $(\mathbf{u}_k)_{k\geq 0}$:

$$\mathbf{u}_k \in, \forall k \geq 0, \lim_{k\to\infty} J(\mathbf{u}_k) = \inf_{\mathbf{v}\in U} J(\mathbf{v})$$

sin excluir en la etapa la posibilidad donde $\inf_{\mathbf{v}\in V} J(\mathbf{v}) = -\infty$. Al estar acotada la sucesión $(\mathbf{u}_k)$ (después de (i)), se demuestra que se puede extraer de ella una sucesión que converge débilmente.

Sea C una constante tal que $\|\mathbf{u}_k\| \leq C$ para todo $k \geq 0$. Se observa para empezar que, si $\mathbf{v}$ es cualquier elemento del espacio V, la sucesión de números reales $\{\langle \mathbf{v}, \mathbf{u}_k\rangle\}_{k\geq 0}$ está acotada ya que $|\langle \mathbf{v}, \mathbf{u}_k\rangle| \leq C\|\mathbf{v}\|$. Suponiendo que el espacio V es separable, sea $(\mathbf{v}_k)_{k\geq 0}$ un conjunto numerable denso. Al estar acotada la sucesión $\{\langle \mathbf{v}_1, \mathbf{u}_k\rangle\}_{k\geq 0}$, se puede extraer una sucesión convergente $\{\langle \mathbf{v}_1, \mathbf{u}_{k_1}\rangle\}_{k_1\geq 0}$; de manera similar, la sucesión $\{\langle \mathbf{v}_1, \mathbf{u}_{k_1}\rangle\}_{k_1\geq 0}$, estando acotada, se puede extraer una sucesión $\{\langle \mathbf{v}_2, \mathbf{u}_{k_2}\rangle\}_{k_2\geq 0}$, convergente, y así sucesivamente.

Se considera la sucesión "diagonal": $(\mathbf{w}_l)_{l\geq 0}$, donde $\mathbf{w}_l \overset{\text{de f}}{=} \mathbf{u}_{l_l}$, Por construcción, cada sucesión $\{\langle \mathbf{v}_k, \mathbf{w}_l\rangle\}_{l\geq 0}$, $k \geq 0$, tiene un límite, que es el límite de la sucesión $\{\langle \mathbf{v}_k, \mathbf{u}_{l_k}\rangle\}_{l_k\geq 0}$. Se demostrará, de hecho que toda sucesión $\{\langle \mathbf{v}, \mathbf{w}_l\rangle\}_{l\geq 0}$, $\mathbf{v} \in V$ tiene un límite: dado cualquier elemento $\mathbf{v} \in V$, y sea $\varepsilon > 0$. Existe un elemento $\mathbf{v}_k$ tal que $\|\mathbf{v} - \mathbf{v}_k\| \leq \frac{\varepsilon}{4C}$, en estas condiciones:

$$\begin{aligned} |\langle \mathbf{v}, \mathbf{w}_l\rangle - \langle \mathbf{v}, \mathbf{w}_l\rangle| &= |\langle \mathbf{v}, \mathbf{w}_l - \mathbf{w}_m\rangle| \\ &\leq |\langle \mathbf{v}_k, \mathbf{w}_l - \mathbf{w}_m\rangle| + |\langle \mathbf{v} - \mathbf{v}_k, \mathbf{w}_l - \mathbf{w}_m\rangle| \\ &\leq |\langle \mathbf{v}_k, \mathbf{w}_l\rangle - \langle \mathbf{v}_k, \mathbf{w}_m\rangle| + \frac{\varepsilon}{2} \end{aligned}$$

Dado que $\|\mathbf{w}_l - \mathbf{w}_m\| \leq \|\mathbf{w}_l\| + \|\mathbf{w}_m\| \leq 2C$. Siendo fijo el elemento $\mathbf{v}_k$, la sucesión $\{\langle \mathbf{v}_k, \mathbf{w}_l\rangle\}_{l\geq 0}$ converge según lo anterior, por lo que es una sucesión de Cauchy. Por tanto, existe un entero $l_0 = l_0(\varepsilon, \mathbf{v}_k)$ tal que:

$$l, k \geq l_0 \Rightarrow |\langle \mathbf{v}_k, \mathbf{w}_l\rangle - \langle \mathbf{v}_k, \mathbf{w}_m\rangle| \leq \frac{\varepsilon}{2}$$

y se establece la afirmación.

Se define una función $f : V \longrightarrow \mathbb{R}$:

$$f(\mathbf{v}) = \lim_{l\to\infty} \langle \mathbf{v}, \mathbf{w}_l\rangle, \forall \mathbf{v} \in V$$

Es una función lineal y continua ya que:

$$|\langle \mathbf{v}, \mathbf{w}_l \rangle| \leq C \|\mathbf{v}\|, \forall l \Rightarrow |f(\mathbf{v})| \leq C \|\mathbf{v}\|$$

Según el *teorema de representación de Riesz*, existe un elemento $\mathbf{u} \in V$ tal que $f(\mathbf{v}) = \langle \mathbf{v}, \mathbf{u} \rangle$ para todo $\mathbf{v} \in V$: por lo tanto, hemos se establece bien la convergencia débil de la sucesión extraída $(\mathbf{w}_l) = (\mathbf{u}_{l_l})$ al elemento $\mathbf{u}$.

(*iii*) Luego se demuestra que el límite "débil" $\mathbf{u}$ de la sucesión extraída $(\mathbf{w}_l)$ pertenece al conjunto U. Se denota por P al operador de proyección asociado con el conjunto convexo U; según el *teorema 1 (2)*:

$$\mathbf{w}_l \in U \Rightarrow \langle P\mathbf{u} - \mathbf{u}, \mathbf{w}_l - P\mathbf{u} \rangle \geq 0, \forall l$$

La convergencia débil de la sucesión $(\mathbf{w}_l)$ hacia el elemento $\mathbf{u}$ implica:

$$0 \leq \lim_{l \to \infty} \langle P\mathbf{u} - \mathbf{u}, \mathbf{w}_l - P\mathbf{u} \rangle = \langle P\mathbf{u} - \mathbf{u}, \mathbf{u} - P\mathbf{u} \rangle = -\|\mathbf{u} - P\mathbf{u}\|^2 \leq 0$$

y por lo tanto $\mathbf{u} \in U$. Se ha establecido así que un conjunto cerrado convexo es "débilmente" cerrado, es decir que le pertenece el límite "débil" de una sucesión de puntos débilmente convergentes de dicho conjunto.

(*iv*) Finalmente, se demuestra que el funcional J satisface:

$$J(\mathbf{v}) = \liminf_{l \to \infty} J(\mathbf{v}_l)$$

para cualquier sucesión $(\mathbf{v}_l)$ que converge débilmente a un elemento $\mathbf{v}$. Suponiendo que la función J es derivable y convexa, se tiene:

$$J(\mathbf{v}) + \langle \nabla J(\mathbf{v}), \mathbf{v}_l - \mathbf{v} \rangle \leq J(\mathbf{v}_l), \forall l$$

y, por definición de convergencia débil:

$$\lim_{l \to \infty} \langle \nabla J(\mathbf{v}), \mathbf{v}_l \rangle = \langle \nabla J(\mathbf{v}), \mathbf{v} \rangle$$

lo que establece la propiedad anunciada; se denomina *semi-continuidad inferior sucesional débil* del funcional J.

(*v*) Ahora es fácil concluir: el límite débil $\mathbf{u} \in U$ de la sucesión extraída $(\mathbf{w}_l)$ de la sucesión minimizadora $(\mathbf{u}_k)$ satisface:

$$J(\mathbf{u}) \leq \liminf_{l \to \infty} J(\mathbf{w}_l) = \lim_{k \to \infty} J(\mathbf{u}_k) = \inf_{\mathbf{v} \in U} J(\mathbf{v})$$

■

○ *Observaciones:*

(*i*) El teorema permanece verdadero en los espacios reflexivos de Banach, de los cuales los espacios de Hilbert (separables o no) son casos especiales: de manera similar, permanece verdadero si se reemplaza la hipótesis de diferenciabilidad de la función J sólo por la continuidad.

(*ii*) El recíproco de la propiedad (*ii*) es verdadero (toda sucesión débilmente convergente está acotada), pero no puede establecerse de forma elemental.

□

En determinados casos particulares, la demostración de la existencia de una solución puede simplificarse notablemente, en particular evitando el recurso a una convergencia débil. Se introduce una definición: dado un espacio de Hilbert V, una función $J : V \longrightarrow \mathbb{R}$ es llamada *funcional cuadrático en* V si es de la forma:

$$J(\mathbf{v}) = \frac{1}{2}a(\mathbf{v},\mathbf{v}) - f(\mathbf{v})$$

donde $a(\mathbf{v},\mathbf{v}) : V \times V \longrightarrow \mathbb{R}$ es una función bilineal, continua, simétrica y $f : V \longrightarrow \mathbb{R}$ es una función lineal continua. Esta definición, naturalmente, generaliza la de un funcional cuadrático sobre $\mathbb{R}^n$ ya que, gracias al teorema de representación de Riesz, existe un operador $A \in \mathcal{L}(V)$ y un elemento $\mathbf{b} \in V$, ambos definidos unívocamente, tales que:

$$a(\mathbf{u},\mathbf{v}) = \langle A\mathbf{u},\mathbf{v}\rangle = \langle \mathbf{u}, A\mathbf{v}\rangle \quad , \quad \forall \mathbf{u},\mathbf{v} \in V$$
$$f(\mathbf{v}) = \langle \mathbf{b},\mathbf{v}\rangle \quad , \quad \forall \mathbf{v} \in V$$

donde $\langle\cdot,\cdot\rangle$ es el producto escalar del espacio V.

El teorema de la proyección y el teorema de la representación de Riesz permiten establecer simplemente un resultado general de existencia para los problemas (P) planteados con tales funcionales. Hay que tener en cuenta que el caso $U = V$ corresponde exactamente a la formulación variacional de los problemas de contorno.

• ***Teorema 6:*** Sea

$$J : \mathbf{v} \in V \longrightarrow J(\mathbf{v}) = \frac{1}{2}a(\mathbf{v},\mathbf{v}) - f(\mathbf{v})$$

Un funcional cuadrático en un espacio de Hilbert V. Además se asume que existe un número α tal que

$$\alpha > 0 \text{ y } a(\mathbf{v},\mathbf{v}) \geq \alpha \|\mathbf{v}\|_V^2, \forall \mathbf{v} \in V$$

Dada un conjunto U de V no vacío, convexo y cerrado, existe uno y solo un elemento $\mathbf{u}$ que satisface:

$$\mathbf{u} \in U \text{ y } J(\mathbf{u}) = \inf_{\mathbf{v}\in U} J(\mathbf{v}) \qquad \text{(P)}$$

Este elemento $\mathbf{u}$ también satisface

$$a(\mathbf{u},\mathbf{v}-\mathbf{u}) \geq f(\mathbf{v}-\mathbf{u}), \forall \mathbf{v} \in U$$

y, a la inversa, si un elemento $\mathbf{u} \in V$ verifica las desigualdades anteriores, esta es la solución del problema (P). Si U es un subespacio vectorial, las desigualdades anteriores se reemplazan por las ecuaciones

$$a(\mathbf{u},\mathbf{v}) = f(\mathbf{v}), \forall \mathbf{v} \in U$$

∘ *Demostración:*

La función bilineal $a(\cdot,\cdot)$ es también es un producto escalar sobre el espacio V, siendo la norma asociada equivalente a la norma $||\cdot||$ asociado con el producto escalar $\langle\cdot,\cdot\rangle$ del espacio V. En efecto, las suposiciones hechas implican:

$$\sqrt{\alpha}\,\|\mathbf{v}\| \leq \sqrt{a(\mathbf{v},\mathbf{v})} \leq \sqrt{\|\alpha\|}\,\|\mathbf{v}\|$$

denotando por $||a||$ a la norma (en el espacio $\mathcal{L}_2(V;\mathbb{R})$) de la función bilineal a.

Dado que la función lineal sigue siendo continua para esta nueva norma, el teorema de representación de Riesz muestra que existe un elemento $\mathbf{c} \in V$ y solo uno tal que

$$f(\mathbf{v}) = a(\mathbf{c}, \mathbf{v}), \forall \mathbf{v} \in V$$

En consecuencia, se puede transformar la expresión del funcional, escribiéndola:

$$J(\mathbf{v}) = \frac{1}{2}a(\mathbf{v}, \mathbf{v}) - a(\mathbf{c}, \mathbf{v}) = \frac{1}{2}a(\mathbf{v}-\mathbf{c}, \mathbf{v}-\mathbf{c}) - \frac{1}{2}a(\mathbf{c}, \mathbf{c})$$

En estas condiciones, resolver el problema (P) equivale a buscar la proyección $\mathbf{u}$ del elemento $\mathbf{c}$ sobre el conjunto U, en el sentido del producto escalar $a(\cdot,\cdot)$. Según el teorema de la proyección, existe uno y solo uno, que establece la existencia y la unicidad de la solución $\mathbf{u}$ del problema (P). Según el mismo teorema, esta solución también se caracteriza por las desigualdades

$$a(\mathbf{u}-\mathbf{c}, \mathbf{v}-\mathbf{u}) \geq 0, \forall \mathbf{v} \in U$$

o por las ecuaciones

$$a(\mathbf{u}-\mathbf{c}, \mathbf{v}) = 0, \forall \mathbf{v} \in U$$

si U es un subespacio vectorial, las relaciones coinciden con las del enunciado ya que $a(\mathbf{c}, \mathbf{v}) = f(\mathbf{v})$, para todo $\mathbf{v} \in V$

■

∘ *Observaciones:*

(*i*) Se ha hecho un uso esencial de la simetría de la función bilineal, por un lado, para concluir que la expresión $a(\cdot,\cdot)$. es un producto escalar, por otro lado, para escribir la nueva expresión del funcional.

(*ii*) Las desigualdades $a(\mathbf{u}, \mathbf{v}-\mathbf{u}) \geq f(\mathbf{v}-\mathbf{u})$ son un caso especial de las desigualdades de Euler $J'(\mathbf{u})(\mathbf{v}-\mathbf{u}) \geq 0$ (teorema 7.4-4) aplicadas al funcional J, de la derivada dada por

$$J'(\mathbf{u})\mathbf{v} = a(\mathbf{u}, \mathbf{v}) - f(\mathbf{v}), \forall \mathbf{v} \in V$$

Se ha hecho una observación similar (y por una buena razón ...) sobre el teorema de la proyección.

□

6.2. Ejemplos de problemas de optimización

La resolución de un nombre lineal en el sentido de mínimos cuadrados es un primer ejemplo de un problema de optimización bajo restricciones, correspondiente a los siguientes datos:

$$U = V = \mathbb{R}^n \; ; \; J : \mathbf{v} \in \mathbb{R}^n \to J(\mathbf{v}) = \frac{1}{2}\|B\mathbf{v} - \mathbf{c}\|_m^2 - \frac{1}{2}\|\mathbf{c}\|_m^2$$

; donde:

$$J(\mathbf{v}) = \frac{1}{2}\left\langle B^t B\mathbf{v}, \mathbf{v}\right\rangle_n - \left\langle B^t\mathbf{c}, \mathbf{v}\right\rangle_m$$

se trata de un problema de *programación cuadrática*, en el sentido aquí entendido, sólo si la matriz simétrica B^tB es definida positiva. Se recuerda que en el apartado 8.1 se estableció la existencia de una solución a este problema en todos los casos, incluido aquel en que la matriz B^tB es únicamente positiva. Cuando la matriz B^tB es definida positiva, la existencia y la unicidad de la solución también se pueden encontrar a partir del *teorema 6*. Una fuente muy grande de problemas de optimización la constituye la *resolución de los problemas en los límites por el método de aproximación variacional*. Este método conduce a encontrar el mínimo de un funcional cuadrático de la forma:

$$\mathcal{J} : \mathbf{v} \in \mathbb{R}^M \to \mathcal{J}(\mathbf{v}) = \frac{1}{2} \langle A\mathbf{v}, \mathbf{v} \rangle - \langle \mathbf{b}, \mathbf{v} \rangle$$

con:

$$A = a(w_j, w_i) \in \mathcal{A}_M(\mathbb{R}) \ , \mathbf{b} = f(w_i) \in \mathbb{R}^M$$

siendo las funciones $w_i, i \leq i \leq M$, las funciones básicas del espacio V_h en las que se busca la solución aproximada $\mathbf{u}_h = \sum_{i=1}^{M} u_i w_i$ y $a(\cdot,\cdot)$ y $f(\cdot)$ interviniendo respectivamente la forma bilineal y la forma lineal en la formulación variacional del problema con los límites considerados. Ya se ha observado que la matriz A es simétrica y definitivamente positiva: por lo tanto, es un segundo ejemplo de un *problema de programación cuadrática sin restricciones*.

Se considera entonces una variante del problema de la membrana, conocido como *la membrana apoyada en un obstáculo* (*figura 4*): se trata de calcular el desplazamiento vertical: $\mathbf{u} : \overline{\Omega} \to \mathbb{R}$ de una membrana elástica de tensión τ, estirado en el borde Γ del abierto $\Omega \to \mathbb{R}^2$ sometido a la acción de una fuerza vertical de densidad $\tau f(\mathbf{x})$ por elemento de superficie, y sujeto a permanecer por encima de un obstáculo representado por una función cono $\mathcal{X} : \overline{\Omega} \to \mathbb{R}$ (de modo que el el problema es posible, se asume la función $\mathcal{X} \leq 0$ en Γ). El área de contacto entre la membrana y el obstáculo no se conoce de antemano.

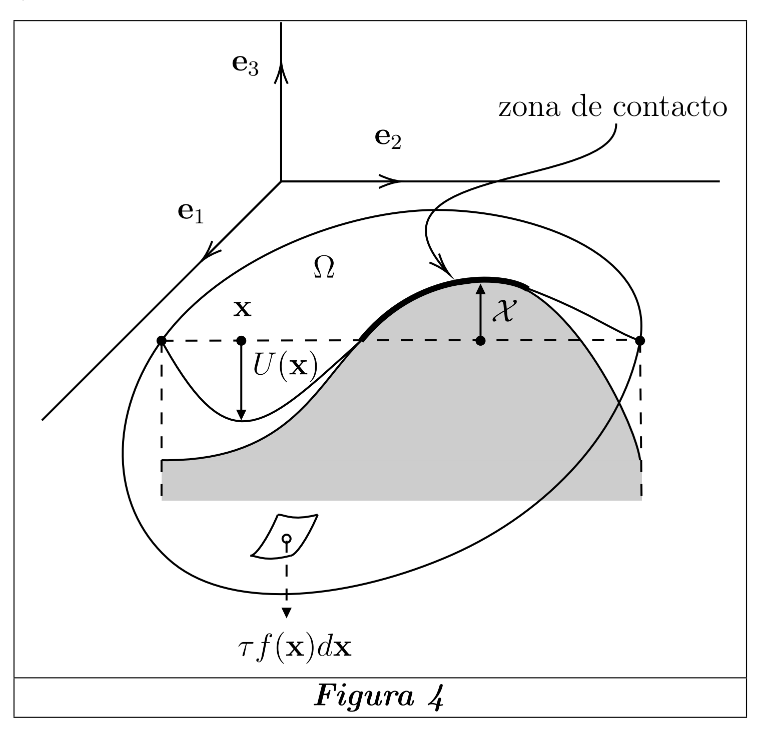


Figura 4

La *formulación variacional* de este problema consiste en buscar el mínimo de la energía de la membrana que es de la forma:

$$J(\mathbf{v}) = \frac{1}{2} a(\mathbf{v}, \mathbf{v}) - f(\mathbf{v})$$

; donde:
$$a(\mathbf{u},\mathbf{v})=\int_\Omega\left(\frac{\partial\mathbf{u}}{\partial x_1}\frac{\partial\mathbf{v}}{\partial x_1}+\frac{\partial\mathbf{u}}{\partial x_2}\frac{\partial\mathbf{v}}{\partial x_2}\right)d\mathbf{x}\ ,\ f(\mathbf{v})=\int_\Omega f\mathbf{v}d\mathbf{x}$$
cuando las funciones $\mathbf{v}$ describen el subconjunto:

$$U=\left\{\mathbf{v}\in V;\mathbf{v}(\mathbf{x})\geq\mathcal{X}(\mathbf{x}).\ ,\text{ para todo }\mathbf{x}\in\overline{\Omega}\right\}$$

un espacio adecuado V de funciones nulas en Γ (este es el espacio $H_0^1(\Omega)$ de Sobolev) Para abordar la solución de este problema, se establece una triangulación del conjunto $\overline{\Omega}$ (supóngase poligonal) y se considera el subespacio $V_h\subset V$ formado por funciones afines en cada triángulo de la triangulación, continua sobre $\overline{\Omega}$ y nula sobre Γ. Recordar que la base "canónica" $(w_i)_{i=1}^M$ de este espacio V_h se elige de tal manera que la función w_I es nula en todos los vértices de la triangulación, excepto en los vértices s_i, donde es igual a 1. En estas condiciones, las componentes de la expansión de una función arbitraria $\mathbf{v}_h\in V_h$ sobre esta base tienen un significado notable ya que:

$$\mathbf{v}_h=\sum_{i=1}^M v_iw_i\Rightarrow v_i=\mathbf{v}_h(\mathbf{s}_i),1\leq i\leq M$$

Por lo tanto, es natural definir el *problema discreto* de la siguiente manera: Encuentre $\mathbf{u}_h$ tal que:

$$\mathbf{u}_h\in U_h=\{\mathbf{u}_h\in V_h;\mathbf{u}_h(\mathbf{s}_i)\geq\mathcal{X}(\mathbf{s}_i),1\leq i\leq M\}\ ,\text{ y } J(\mathbf{u}_h)=\inf_{\mathbf{v}_h\in U_h}J(\mathbf{v}_h)$$

Notar que el conjunto U_h, generalmente no está contenido en el conjunto U. Por lo tanto, se busca el mínimo de la funcional cuadrática a lo largo de la cual:

$$\mathcal{J}:\mathbf{v}\in\mathbb{R}^M\rightarrow\mathcal{J}(\mathbf{v})=\frac{1}{2}\langle A\mathbf{v},\mathbf{v}\rangle-\langle\mathbf{b},\mathbf{v}\rangle$$

, con:
$$A=a(w_j,w_i)\in\mathcal{A}_M(\mathbb{R})\ ,\ \mathbf{b}=f(w_i)\in\mathbb{R}^M$$
cuando el vector $\mathbf{v}$ describe el conjunto:

$$\mathcal{U}=\left\{\mathbf{v}=(v_i)\in\mathbb{R}^M;v_i\geq\mathcal{X}(s_i),1\leq i\leq M\right\}$$

Es por tanto un ejemplo de un *problema de programación cuadrática con restricciones-desigualdades afines*. El conjunto siendo $\mathcal{U}$ no vacío, cerrado, convexo (y no acotado), la existencia de una solución del problema discreto resulta tanto del *teorema 4* como del *teorema 6*. Notar que se puede escribir el problema discreto en la forma equivalente: Encontrar $\mathbf{u}_h$ tal que:
$$\mathbf{u}_h\in U_h\text{ y } a(\mathbf{u}_h,\mathbf{v}_h-\mathbf{u}_h)\geq f(\mathbf{v}_h-\mathbf{u}_h)\ ,\text{ para todo }\mathbf{u}_h\in U_h$$
Estando definido el espacio V como anteriormente, el *problema de la torsión-elastoplástica de una barra cilíndrica* lleva a buscar el mínimo del mismo funcional:

$$J(\mathbf{v})=\frac{1}{2}a(\mathbf{v},\mathbf{v})-f(\mathbf{v})=\frac{1}{2}\int_\Omega\left(\left(\frac{\partial\mathbf{v}}{\partial x_1}\right)^2+\left(\frac{\partial\mathbf{v}}{\partial x_2}\right)^2\right)d\mathbf{x}-\int_\Omega f\mathbf{v}d\mathbf{x}$$

cuando las funciones $\mathbf{v}$ describen el subconjunto:

$$U=\left\{\mathbf{v}\in V;\|\nabla\mathbf{v}(\mathbf{x})\|\leq1\text{ para casi todo }\mathbf{x}\in\overline{\Omega}\right\}$$

planteando:

$$\|\nabla \mathbf{v}(\mathbf{x})\| = \left(\left(\frac{\partial \mathbf{v}}{\partial x_1}(\mathbf{x})\right)^2 + \left(\frac{\partial \mathbf{v}}{\partial x_2}(\mathbf{x})\right)^2\right)^{\frac{1}{2}}$$

El *problema discreto* asociado al espacio V, de elementos finitos introducidos más frecuentado consiste en buscar el mínimo del funcional J cuando la función $\mathbf{v}_h$ describe el conjunto:

$$U_h = \left\{\mathbf{v}_h \in V_h; \|\nabla \mathbf{v}_h(\mathbf{x})\| \leq 1 \text{ para todo } \mathbf{x} \in \mathring{T}, T \in \mathcal{T}_h\right\}$$

donde $\mathring{T}$ denota el interior de cada uno de los triángulos T de la triangulación $\mathcal{T}_h$. Es fácil ver que el conjunto U_h es no vacío, cerrado y convexo ya que:

$$\mathbf{v}, \mathbf{w} \in U \text{ y } \theta \in [0,1] \Rightarrow \|\nabla\left(\theta \mathbf{v} + (1-\theta)\mathbf{w}\right)(\mathbf{x})\| \leq \theta \|\nabla \mathbf{v}(\mathbf{x})\| + (1-\theta)\|\nabla \mathbf{w}(\mathbf{x})\| \leq 1$$

por lo que el problema de optimización asociado aún admite una solución y sólo una , que se puede caracterizar de manera equivalente por desigualdades variacionales. Notar que el conjunto $U_h = U \cap V_h$ está esta vez contenido en el conjunto U Sea T un triángulo de la triangulación $\mathcal{T}_h$, de vértices $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3$ (para fijar las ideas). La restricción de cualquier función $\mathbf{v} \in V_h$ al triángulo T se escribe:

$$\mathbf{v}|_T = \sum_{i=1}^{3} v_i w_i|_T \text{ , con } v_i = \mathbf{v}(\mathbf{s}_i)$$

Dado que las funciones base w son afines, sus primeras derivadas son constantes, por lo que la desigualdad $\|\nabla(\mathbf{v}|_T)\| \leq 1$ toma la forma:

$$\left(\sum_{i=1}^{3} \alpha_i v_i\right)^2 + \left(\sum_{i=1}^{3} \beta_i v_i\right)^2 \leq 1$$

siendo las constantes $\alpha_i = \frac{\partial(w_i|_T)}{\partial x_1}$ y $\beta_i = \frac{\partial(w_i|_T)}{\partial x_2}$ funciones conocidas de las coordenadas de los vértices $\mathbf{s}_i$. Se está por tanto en presencia de un *problema de programación cuadrática con m desigualdades cuadráticas restringidas* (m = número de triángulos de la triangulación $\mathcal{T}_h$).

6.3. Ejercicios

Ejercicio 6.1: Demuestre que una sucesión $\{\mathbf{u}_k\}_{k=0}^{\infty}$ de elementos de un espacio de Hilbert V converge (en el sentido de la norma) hacia un elemento $\mathbf{u} \in V$ si, y solo si, converge débilmente hacia este mismo elemento y $\lim_{k\to\infty} \|\mathbf{u}_k\| = \|\mathbf{u}\|$.

○ *Solución*

"$\Rightarrow$"

$\mathbf{u}_k$ converge a $\mathbf{u}$ en sentido de la norma , es decir que dado un $\varepsilon > 0, \exists N \in \mathbb{N}$ tal que: $\|\mathbf{u}_k - \mathbf{u}\| < \varepsilon, k \geq N$.

Sea $\mathbf{v} \in V$ y $\varepsilon > 0$:

$$\begin{aligned} |\langle \mathbf{u}_k, \mathbf{v}\rangle - \langle \mathbf{u}, \mathbf{v}\rangle| &= |\langle \mathbf{u}_k - \mathbf{u}, \mathbf{v}\rangle| \\ &\leq \|\mathbf{u}_k - \mathbf{u}\| \|\mathbf{v}\| \text{ , por desigualdad de Cauchy-Schwarz} \\ &= 0 < \varepsilon, k \to \infty \end{aligned}$$

; por lo que $\langle \mathbf{u}_k, \mathbf{v} \rangle \to \langle \mathbf{u}, \mathbf{v} \rangle, \forall \mathbf{v} \in V, k \to \infty$ es decir que $\mathbf{u}_k$ converge débilmente a $\mathbf{u}$.

Ahora se considera: $|\|\mathbf{u}_k\| - \|\mathbf{u}\||$:

$$\begin{aligned} |\|\mathbf{u}_k\| - \|\mathbf{u}\|| &\leq |\|\mathbf{u}_k - \mathbf{u}\|| \text{ , por propiedades de norma} \\ &< \varepsilon, k \to \infty \end{aligned}$$

; por lo que $\|\mathbf{u}_k\| \to \|\mathbf{u}\|, k \to \infty$.

"$\Leftarrow$"

Se tiene que $\mathbf{u}_k$ converge débilmente a $\mathbf{u}$ y $\|\mathbf{u}_k\| \to \|\mathbf{u}\|, k \to \infty$, luego se considera:

$$\begin{aligned} \|\mathbf{u}_k - \mathbf{u}\|^2 &= \langle \mathbf{u}_k - \mathbf{u}, \mathbf{u}_k - \mathbf{u} \rangle \\ &= \|\mathbf{u}_k\|^2 - \langle \mathbf{u}_k, \mathbf{u} \rangle - \langle \mathbf{u}, \mathbf{u}_k \rangle + \|\mathbf{u}\|^2 \\ &= \|\mathbf{u}\|^2 - \langle \mathbf{u}, \mathbf{u} \rangle - \langle \mathbf{u}, \mathbf{u} \rangle + \|\mathbf{u}\|^2, k \to \infty \\ &= 0 \end{aligned}$$

; por lo cual $\mathbf{u}_k$ converge a $\mathbf{u}$ en sentido de la norma.

$\boxtimes$

Ejercicio 6.2: Consideramos un funcional cuadrático:

$$J(\mathbf{v}) = \frac{1}{2} \langle A\mathbf{v}, \mathbf{v} \rangle - \langle \mathbf{b}, \mathbf{v} \rangle, A \in \mathcal{A}_n(\mathbb{R}), \mathbf{b} \in \mathbb{R}^n,$$

la matriz simétrica A es definida positiva, y el conjunto (se supone que no está vacío):

$$U = \{\mathbf{v} \in \mathbb{R}^n / C\mathbf{v} = \mathbf{d}\}, C \in \mathcal{A}_{m,n}(\mathbb{R}), \mathbf{d} \in \mathbb{R}^m.$$

(i) Muestre que el problema asociado (P) tiene una solución y solo una.

(ii) Demuestre que un vector $\mathbf{u} \in \mathbb{R}^n$ es la solución del problema si, y solo si, existe un vector $\lambda \in \mathbb{R}^m$ tal que:

$$\begin{cases} A\mathbf{u} + C^t\lambda &= \mathbf{b} \\ C\mathbf{u} &= \mathbf{d} \end{cases}$$

(iii) Se supone que el rango de la matriz C es m. Exprese la solución $\mathbf{u}$ en función de los datos $A, \mathbf{b}, C, \mathbf{d}$.

∘ *Solución*

(i) Se consideran los siguientes elementos:

• ***Definición (Funcional cuadrático en $\mathbb{R}^n$):*** Un *funcional cuadrático sobre* $\mathbb{R}^n$ tiene la forma:

$$J(\mathbf{v}) = \frac{1}{2} \langle A\mathbf{v}, \mathbf{v} \rangle - \langle \mathbf{b}, \mathbf{v} \rangle$$

; donde $A \in \mathcal{A}_n(\mathbb{R})$ es una matriz simétrica dada y $\mathbf{b} \in \mathbb{R}^n$ un vector cualquiera. Tal que es diferenciable en $\mathbb{R}^n$ y:

$$\nabla J(\mathbf{u}) = A\mathbf{u} - \mathbf{b}$$

• ***Definición (Funcional cuadrático):*** Sea V un espacio de Hilbert, un funcional $J : V \longrightarrow \mathbb{R}$ se llama *funcional cuadrático sobre* V si tiene la forma:

$$J(\mathbf{v}) = \frac{1}{2}a(\mathbf{v}, \mathbf{v}) - f(\mathbf{v})$$

; donde $a : V \times V \longrightarrow \mathbb{R}$ es bilineal, continua y simétrica y f es función lineal y continua.

En el problema planteado se tiene que J es un funcional cuadrático sobre V por hipótesis, luego se define una función $h : V \longrightarrow \mathbb{R}, h(\mathbf{v}) = a(\mathbf{v}, \mathbf{v})$ y se considera el conjunto $S = \{\mathbf{v} \in V / \|\mathbf{v}\| = 1\}$, notar que es cerrado y acotado (compacto en V) y por el teorema de Weierstrass alcanza un mínimo en S dado que h es continua, luego:

$$\alpha = h(\mathbf{u}) = \min_{\mathbf{v} \in S} h(\mathbf{v}) = \min_{\mathbf{v} \in S} a(\mathbf{v}, \mathbf{v}) = \min_{\mathbf{v} \in S} \langle A\mathbf{v}, \mathbf{v} \rangle$$

, se tiene que $\alpha > 0$ porque $\mathbf{u} \in S$ y por propieades del producto escalar $< \cdot, \cdot >$, luego para $\mathbf{w} \in V$ con $\mathbf{w} \neq 0$:

$$h\left(\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) = a\left(\frac{\mathbf{w}}{\|\mathbf{w}\|}, \frac{\mathbf{w}}{\|\mathbf{w}\|}\right) \geq \alpha \quad \Rightarrow \quad a(\mathbf{w}, \mathbf{w}) \geq \alpha \|\mathbf{w}\|^2$$

; notar que a es bilineal y continua (esto viene de que J es funcional cuadrático); en el caso de $\mathbf{w} = 0$ es claro que $a(\mathbf{w}, \mathbf{w}) = 0 = \|\mathbf{w}\|^2$, por lo cual $\exists \alpha > 0$ tal que:

$$\alpha > 0 \text{ y } a(\mathbf{v}, \mathbf{v}) \geq \alpha \|\mathbf{v}\|_V^2, \forall \mathbf{v} \in V$$

, se sabe que U es no vacío por hipótesis, ahora hay que notar que es convexo y cerrado.

$$U = \{\mathbf{v} \in \mathbb{R}^n / C\mathbf{v} = \mathbf{d}\}, C \in \mathcal{A}_{m,n}(\mathbb{R}), \mathbf{d} \in \mathbb{R}^m,$$

$$\begin{aligned} \text{Sean } \mathbf{u}, \mathbf{v} \in U \text{ ;ahora se considera:} \quad \mathbf{w} &= \lambda\mathbf{u} - (1-\lambda)\mathbf{v}, \lambda \in [0,1] \\ \Rightarrow \quad C\mathbf{w} &= \lambda C\mathbf{u} - (1-\lambda)C\mathbf{v} \\ &= \lambda\mathbf{d} - (1-\lambda)\mathbf{d} \\ &= \mathbf{d} \end{aligned}$$

Por lo que $\mathbf{w} \in U$ y U es convexa.

Sea $\{\mathbf{u}_k\}_{k=0}^{\infty} \in U$ tal que $\mathbf{u}_k \to \mathbf{u}$, luego $C\mathbf{u}_k$ converge a $C\mathbf{u}$ y también a $\mathbf{d}$, por lo que $C\mathbf{u} = \mathbf{d}$ es decir que $\mathbf{u} \in U$ y se tiene que U es cerrado.

Ahora usando el *teorema 6* se concluye que el problema (P) asociado a J tiene solución única en U

(ii) " $\Leftarrow$ "

Se considera el siguiente resultado:

• ***Teorema:*** Sea U un subconjunto convexo de un espacio normado V:

Sea $J : \Omega \subset V \longrightarrow \mathbb{R}$ una función convexa definida en un Ω abierto de V que contiene a U, diferenciable en un punto $\mathbf{u} \in U$. Entonces la función J admite un mínimo en U con respecto al conjunto U sí y sólo sí:

$$J'(\mathbf{u})(\mathbf{v} - \mathbf{u}) \geq 0 \text{ ,para todo } \mathbf{v} \in U$$

Sean $\mathbf{u}, \mathbf{v} \in U$ $(C\mathbf{u} = \mathbf{d})$ y se supone que existe un $\lambda \in \mathbb{R}^n$ tal que: $A\mathbf{u} + C^t\lambda = \mathbf{b}$, y tomando en cuenta la *definición 1* se tiene:

$$\begin{aligned}
J'(\mathbf{u})(\mathbf{v}-\mathbf{u}) &= \langle \nabla J(\mathbf{u}), \mathbf{v}-\mathbf{u} \rangle \\
&= \langle A\mathbf{u}-\mathbf{b}, \mathbf{v}-\mathbf{u} \rangle \\
&= -\left\langle C^t\lambda, \mathbf{v}-\mathbf{u} \right\rangle \\
&= -\left(C^t\lambda\right)^t(\mathbf{v}-\mathbf{u}) \\
&= -\lambda^t C(\mathbf{v}-\mathbf{u}) \\
&= -\lambda^t(C\mathbf{v}-C\mathbf{u}) \\
&= -\langle \lambda, C\mathbf{v}-C\mathbf{u} \rangle \\
&= -\langle \lambda, \mathbf{b}-\mathbf{b} \rangle \\
&= 0
\end{aligned}$$

; luego aplicando el teorema anterior se tiene que $\mathbf{u}$ es solución del problema (P) en U

"$\Rightarrow$"

• ***Teorema:*** Sea U un subconjunto convexo de un espacio normado V:

Sea Ω un conjunto abierto de $\mathbb{R}^n$,y $\varphi_i : \Omega \longrightarrow \mathbb{R}, 1 \leq i \leq m$ funciones de clase $\mathcal{C}^1$ sobre Ω y sea $\mathbf{u}$ un punto del conjunto

$$U = \{\mathbf{v} \in \Omega / \varphi_i(\mathbf{v}) = 0, 1 \leq i \leq m\} \subset \Omega$$

en el que las derivadas de $\varphi_i'(\mathbf{u}) \in \mathcal{L}(\mathbb{R}^n; \mathbb{R})$ (son funcionales lineales), $1 \leq i \leq m$, son linealmente independientes.

Sea $J : \Omega \longrightarrow \mathbb{R}$ una función diferenciable en U. si la función admite un $\mathbf{u}$ extremo relativo respecto al conjunto U, existen m números $\lambda_i(\mathbf{u}), 1 \leq i \leq m$ definidos de forma única como:

$$J'(\mathbf{u}) + \lambda_1(\mathbf{u})\varphi_1'(\mathbf{u}) + \lambda_2(\mathbf{u})\varphi_2'(\mathbf{u}) + \cdots + \lambda_m(\mathbf{u})\varphi_m'(\mathbf{u}) = 0$$

Se asume $m < n$. Luego se considera la función $\varphi : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ tal que $\varphi(\mathbf{v}) = C\mathbf{v} - \mathbf{d}$ con lo cual se definen m funciones reales tales que: $\varphi_i(\mathbf{v}) = (C\mathbf{v} - \mathbf{d})_i$, para $1 \leq i \leq m$ y se tiene:

$$U = \{\mathbf{v} \in \mathbb{R}^n | \varphi_i(\mathbf{v}) = (C\mathbf{v} - \mathbf{d})_i = 0, 1 \leq i \leq m\}$$

Notar que:

$$\begin{aligned}
\begin{pmatrix} \partial_1\varphi_1(\mathbf{u}) & \cdots & \partial_1\varphi_m(\mathbf{u}) \\ \vdots & \ddots & \vdots \\ \partial_n\varphi_1(\mathbf{u}) & \cdots & \partial_n\varphi_m(\mathbf{u}) \end{pmatrix} \begin{pmatrix} \lambda_1(\mathbf{u}) \\ \vdots \\ \lambda_m(\mathbf{u}) \end{pmatrix} &= \begin{pmatrix} c_{11}(\mathbf{u}) & \cdots & c_{m1}(\mathbf{u}) \\ \vdots & \ddots & \vdots \\ c_{1n}(\mathbf{u}) & \cdots & c_{mn}(\mathbf{u}) \end{pmatrix} \begin{pmatrix} \lambda_1(\mathbf{u}) \\ \vdots \\ \lambda_m(\mathbf{u}) \end{pmatrix} \\
&= C^t\lambda(\mathbf{u})
\end{aligned}$$

Se deduce del anterior que una condición necesaria para que la función J admita en un punto $\mathbf{u} \in U$ un extremo relativo respecto de U es la existencia de una solución

$(\mathbf{u}, \lambda) \in \mathbb{R}^{n+m}$ de la ecuación matricial:

$$\begin{pmatrix} A & C^t \\ C & 0 \end{pmatrix}\begin{pmatrix} \mathbf{u} \\ \lambda \end{pmatrix} = \begin{pmatrix} A\mathbf{u} + C^t\lambda \\ C\mathbf{u} \end{pmatrix}$$
$$= \begin{pmatrix} \nabla J(\mathbf{u}) + \mathbf{b} + C^t\lambda \\ \mathbf{d} \end{pmatrix}$$
$$= \begin{pmatrix} \nabla J(\mathbf{u}) + C^t\lambda + \mathbf{b} \\ \mathbf{d} \end{pmatrix}$$
$$= \begin{pmatrix} \mathbf{b} \\ \mathbf{d} \end{pmatrix}$$

lo que implica la existencia de solución del sistema lineal:

$$\begin{cases} A\mathbf{u} + C^t\lambda & = \mathbf{b} \\ C\mathbf{u} & = \mathbf{d} \end{cases}$$

(iii) Notar que la matriz A es invertible por ser definida positiva, luego para $\mu \in \mathbb{R}^m$ se tiene que:

$$\mu^t C A^{-1} C^t \mu = C^t \mu A^{-1} C^t \mu = 0 \Rightarrow C^t\mu = 0 \Rightarrow \mu = 0$$

, además $ker\,(C^t) = \{\mathbf{0}\}$ $(C\mathbf{v} = \mathbf{d})$ por lo que $CA^{-1}C^t$ es invertible, luego:

$$\begin{aligned} C^t\lambda = \mathbf{b} - A\mathbf{u} \Rightarrow A^{-1}C^t\lambda &= A^{-1}\mathbf{b} - \mathbf{u} \\ \Rightarrow CA^{-1}C^t\lambda &= CA^{-1}\mathbf{b} - C\mathbf{u} \\ \Rightarrow \lambda &= \left(CA^{-1}C^t\lambda\right)^{-1}\left(CA^{-1}\mathbf{b} - \mathbf{d}\right) \\ &= \left(CA^{-1}C^t\lambda\right)^{-1} CA^{-1}\mathbf{b} - \left(CA^{-1}C^t\lambda\right)^{-1}\mathbf{d} \end{aligned}$$

Luego: $A\mathbf{u} = \mathbf{b} - C^t\lambda \Rightarrow \mathbf{u} = A^{-1}\mathbf{b} - A^{-1}C^t\lambda$

$$\Rightarrow \mathbf{u} \;=\; A^{-1}\mathbf{b} - A^{-1}C^t\left(CA^{-1}C^t\right)^{-1}CA^{-1}\mathbf{b} + A^{-1}C^t\left(CA^{-1}C^t\right)^{-1}\mathbf{d}$$

⊠

Ejercicio 6.3: Se considera el problema de las *superficies mínimas*: Entre todas las superficies que se apoyan en un contorno del espacio $\mathbb{R}^3$, se encuentra la superficie mínima. El contorno representado por la función $\mathbf{u}_0(\mathbf{x})$, $\mathbf{x} \in \Gamma$ donde Γ es el borde de un abierto Ω del plano $\mathbb{R}^2$, se trata de hallar el mínimo de la función:

$$J(\mathbf{v}) = \int_\Omega \sqrt{1 + \parallel \nabla\mathbf{v} \parallel^2} d\mathbf{x}$$

Cuando $\mathbf{v}$ describe un conjunto de funciones definidas en $\overline{\Omega}$, suficientemente regulares e iguales a la función $\mathbf{u}_0$ en Γ. Dada una triangulación del conjunto Ω, se considera el espacio V_h formado por las funciones afines en cada triangulo de la triangulación y continuas en Γ. Notar que $\sum_h$ es el conjunto de vértices de la triangulación que se encuentran en Γ y se define el conjunto:

$$U_h = \{\mathbf{v}_h \in V_h | \mathbf{v}_h(\mathbf{s}) = \mathbf{u}_0(\mathbf{s}), \forall \mathbf{s} \in \textstyle\sum_h\}$$

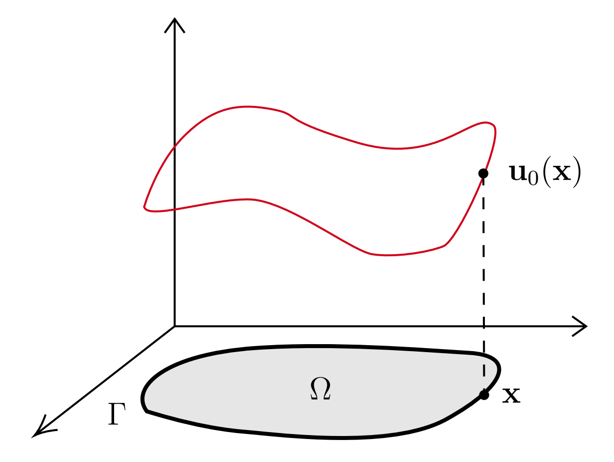
Demostrar que el problema discreto: Encontrar $\mathbf{u}_h$ tal que el problema:

$$\mathbf{u}_h \in U_h \text{ y } J(\mathbf{u}_h) = \inf_{\mathbf{v}_h \in U_h} J(\mathbf{v}_h)$$

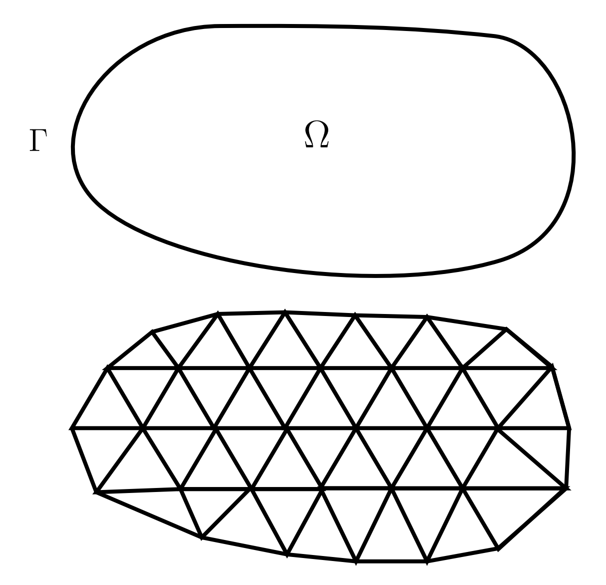
tenga una solución y sólo una. Hay que tener en cuenta que este es un ejemplo de función no cuadrática $J : V_h \longrightarrow \mathbb{R}$

∘ *Solución*

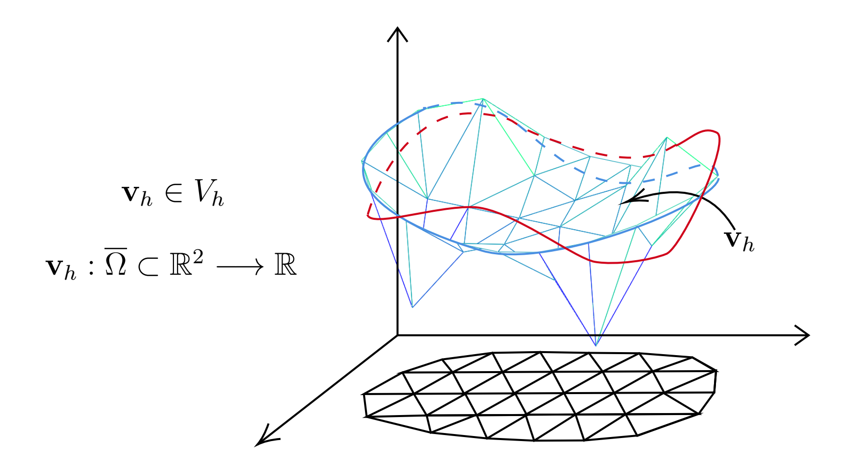
Primero hay que discretizar (nunca mejor dicho) los elementos del problema. Se considera el contorno generado por $\mathbf{u}_0(\mathbf{x})$, $\mathbf{x} \in \Gamma$ donde Γ es el contorno de un abierto $\Omega \in \mathbb{R}^2$.



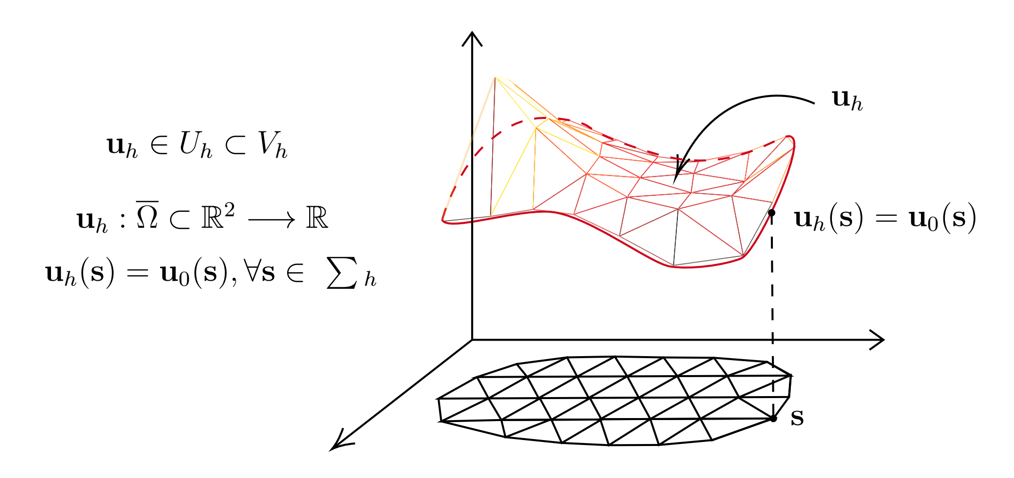
Luego se da *triangulación* definida por un real positivo h (es el mayor valor que puede tomar el diámetro de cada triángulo de la triangulación):



Ahora se considera el conjunto V_h de funciones en $\overline{\Omega}$ formadas por funciones afines (lineales) en cada triángulo y tales que sean continuas en Γ:



Además se define un subconjunto U_h de V_h que cumple que para $\mathbf{u}_h \in U_h$: $\mathbf{u}_h(\mathbf{s}) = \mathbf{u}_0(\mathbf{s}), \forall \mathbf{s} \in \sum_h$ donde $\sum_h$ son todos los vértices de la triangulación que están en Γ:



Ahora hay que demostrar que existe un $\mathbf{u}_h$ tal que:

$$\mathbf{u}_h \in U_h \text{ y } J(\mathbf{u}_h) = \inf_{\mathbf{v}_h \in U_h} J(\mathbf{v}_h) \tag{P}$$

, para ello se tomará en cuenta lo siguiente:

• ***Definición:*** Una función J con valores reales definidos en un espacio vectorial normado V se dice *coercitiva* si:

$$\lim_{\|\mathbf{v}\|_V \to \infty} J(\mathbf{v}) = +\infty$$

El conjunto U_h es no vacío (se puede considerar la función nula para todo vértice que no sea parte del contorno). Sea $\overline{\mathbf{u}}_h$ cualquier función de U_h (por ejemplo la función definida por $\overline{\mathbf{u}}_h(\mathbf{s}) = \mathbf{u}_0(\mathbf{s})$ para todo $\mathbf{s} \in \sum_h$ y $\overline{\mathbf{u}}_h(\mathbf{s}) = 0$ para todo vértice interior de la triangulación). Se denota por $\widetilde{U}_h$ el subespacio vectorial de V_h definido por:

$$\widetilde{U}_h = \{\mathbf{v}_h \in V_h | \mathbf{v}_h = 0 \text{ en } \Gamma\}$$

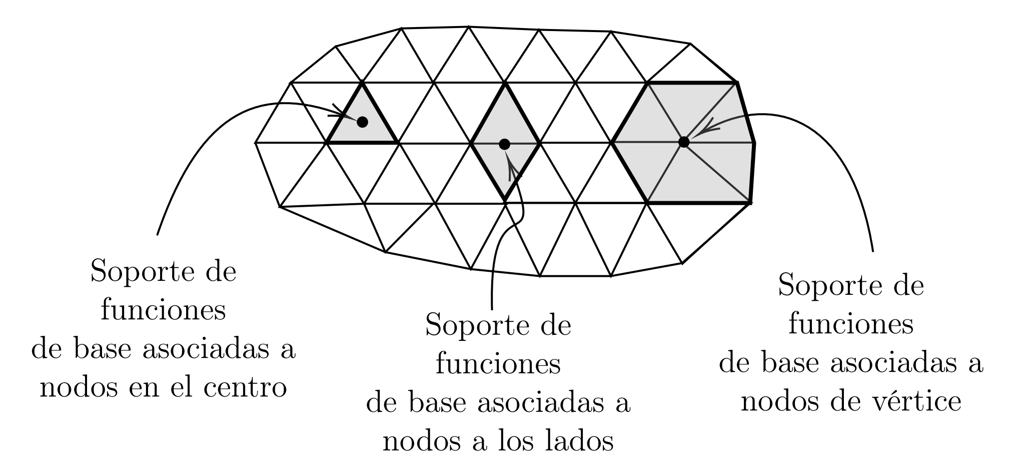
Así cualquier función $\mathbf{v}_h \in U_h$ se escribe $\mathbf{v}_h = \overline{\mathbf{u}}_h + \widetilde{\mathbf{v}}_h$ con $\widetilde{\mathbf{v}}_h \in \widetilde{U}_h$ y el problema de minimización propuesto se reduce a encontrar $\widetilde{\mathbf{u}}_h \in \widetilde{U}_h$ tal que:

$$\widetilde{\mathbf{u}}_h \in \widetilde{U}_h \text{ y } \widetilde{J}\left(\widetilde{\mathbf{u}}_h\right) = \inf_{\widetilde{\mathbf{v}}_h \in \widetilde{U}_h} \widetilde{J}\left(\widetilde{\mathbf{v}}_h\right) \tag{P}$$

donde $\widetilde{J} : \widetilde{U}_h \subset V_h \longrightarrow \mathbb{R}$ está definida de esta forma:

$$\widetilde{J}\left(\widetilde{\mathbf{v}}_h\right) = J\left(\widetilde{\mathbf{v}}_h + \overline{\mathbf{u}}_h\right) = \int_\Omega \sqrt{1 + \left\|\nabla\left(\widetilde{\mathbf{v}}_h + \overline{\mathbf{u}}_h\right)\right\|^2} d\mathbf{x}$$

Para notar la cerradura de $\widetilde{U}_h$ hay que observar cómo se construyen las funciones de este conjunto, hay varias formas de construirlas por medio de las funciones de base a partir de los vértices de la triangulación (*nodos*):

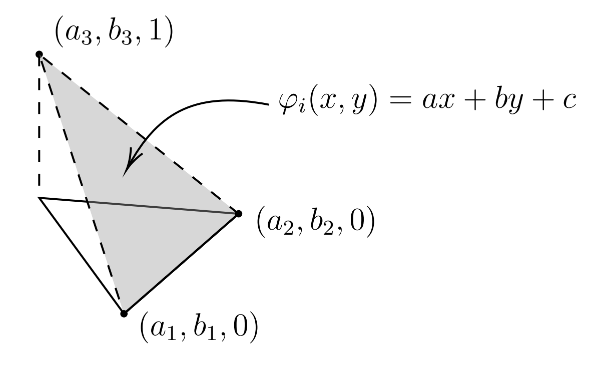


En todo caso las funciones base están definidas en términos de los nodos, suponiendo que estas son del tercer tipo (funciones de base asociadas a nodos de vértice) se tiene que las funciones de V_h tienen la siguiente forma:

$$\mathbf{u}_h = \sum_{i=1}^{N_h} n_i \varphi_i(\mathbf{s}) = \sum_{i=1}^{N_h} n_i \varphi_i(x, y)$$

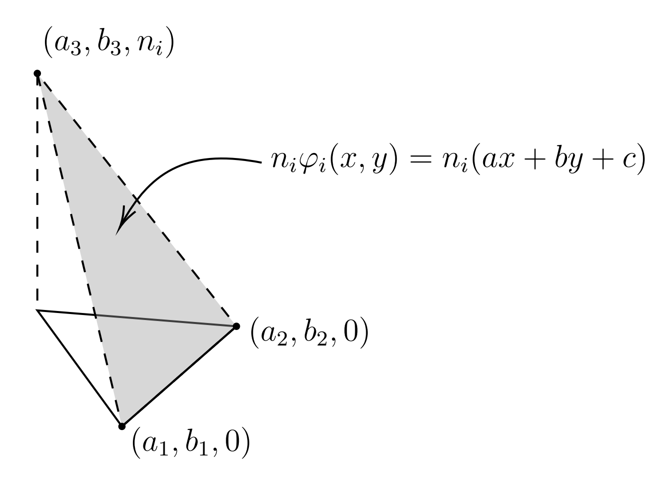
; donde: $N_h :=$ número de nodos considerados
$\varphi_i :=$ función base asociada al nodo i
$n_i :=$ función que define la altura de cada función base según el nodo i

Luego cada función φ_i tiene la forma:



; donde a, b y c dependen de a_1, a_2, a_3, b_1, b_2 y b_3.

Y cada función n_i son constantes que dependen de cada $i-$nodo por lo cual:



Así una sucesión de funciones de $\widetilde{U}_h \subset V_h$: $\{\widetilde{\mathbf{u}}_{h,k}\}_{k=1}^{\infty}$ tiene términos de esta forma:

$$\widetilde{\mathbf{u}}_{h,k} = \sum_{i=1}^{N_h} n_{i,k}\varphi_i(\mathbf{s})$$

, así está claro que cada término de la sucesión de funciones está definido por cada $n_{i,k}$ que son constantes, notar que $n_{i,k} = 0$ si el índice i corresponde a algún nodo que está en Γ, luego si la sucesión es convergente se tiene que converge en $\widetilde{U}_h$ dado que para los nodos en Γ, la sucesión $n_{i,k}$ es constante nula y en los demás nodos deberá existir algún real n^* tal que $n_{i,k} \to n^*$ por lo que se cumple la cerradura.

Luego el funcional $\widetilde{J}$ es coercitivo, porque:

$$\begin{array}{rcl} \widetilde{J}\left(\widetilde{\mathbf{v}}_h\right) & = & \int_\Omega \sqrt{1 + \left\|\nabla\left(\widetilde{\mathbf{v}}_h + \overline{\mathbf{u}}_h\right)\right\|^2} d\mathbf{x} \\ & \geq & \int_\Omega \left\|\nabla\left(\widetilde{\mathbf{v}}_h + \overline{\mathbf{u}}_h\right)\right\| d\mathbf{x} \\ & \geq & \int_\Omega \left\|\nabla\widetilde{\mathbf{v}}_h\right\| d\mathbf{x} - \int_\Omega \|\nabla\overline{\mathbf{u}}_h\| d\mathbf{x} \end{array}$$

, así cuando $\|\widetilde{\mathbf{v}}_h\| \to \infty$ entonces $\widetilde{J}\left(\widetilde{\mathbf{v}}_h\right) \to \infty$, por otro lado $\widetilde{\mathbf{v}}_h \to \int_\Omega \left\|\nabla\widetilde{\mathbf{v}}_h\right\| d\mathbf{x}$ es una norma sobre $\widetilde{U}_h$ (no es difícil demostrarlo mediante propiedades de la norma $||\cdot||$ y de la integral). Se deduce del *teorema 4* que el problema de minimización admite al menos una solución $\widetilde{\mathbf{u}}_h \in \widetilde{U}_h$, por lo tanto existe un $\mathbf{u}_h = \overline{\mathbf{u}}_h + \widetilde{\mathbf{u}}_h \in U_h$ que es solución del problema (P).

Ahora para verificar que esta solución es única se consideran los teoremas:

• ***Teorema:*** Sea $J : U \subset \Omega \longrightarrow \mathbb{R}$ una función dos veces derivable en un abierto Ω de un espacio vectorial normado V, y U un subconjunto convexo de Ω

Si

$$J''(\mathbf{u})(\mathbf{v} - \mathbf{u})(\mathbf{v} - \mathbf{u}) > 0,\ \forall \mathbf{u}, \mathbf{v} \in U, \mathbf{u} \neq \mathbf{v}$$

, la función J es estrictamente convexa en U.

• ***Teorema:*** Sea U un subconjunto convexo de un espacio normado V:

Una función $J : U \subset \Omega \longrightarrow \mathbb{R}$ estrictamente convexa admite como máximo un mínimo, y es un mínimo estricto

Se observa que $\widetilde{J}$ es dos veces derivable, luego se analiza el producto vectorial $\widetilde{J}''(\widetilde{\mathbf{u}}_h)(\widetilde{\mathbf{w}}_h, \widetilde{\mathbf{w}}_h)$ (notar que el operador $\widetilde{J}''(\widetilde{\mathbf{u}}_h)$ se aplica sobre el espacio $V_h \times V_h$), luego:

$$
\begin{aligned}
\widetilde{J}''(\widetilde{\mathbf{u}}_h)(\widetilde{\mathbf{w}}_h, \widetilde{\mathbf{w}}_h) &= \left\langle \nabla^2 \widetilde{J}(\widetilde{\mathbf{u}}_h)\,\widetilde{\mathbf{w}}_h, \widetilde{\mathbf{w}}_h \right\rangle, \text{ Ciarlet p.145} \\
&= \widetilde{\mathbf{w}}_h^t \nabla^2 \widetilde{J}(\widetilde{\mathbf{u}}_h)\,\widetilde{\mathbf{w}}_h \\
&= \widetilde{\mathbf{w}}_h^t \begin{pmatrix} \frac{\partial^2 \widetilde{J}(\widetilde{\mathbf{u}}_h)}{\partial x^2} & \frac{\partial^2 \widetilde{J}(\widetilde{\mathbf{u}}_h)}{\partial x \partial y} \\ \frac{\partial^2 \widetilde{J}(\widetilde{\mathbf{u}}_h)}{\partial y \partial x} & \frac{\partial^2 \widetilde{J}(\widetilde{\mathbf{u}}_h)}{\partial y^2} \end{pmatrix} \widetilde{\mathbf{w}}_h \\
&= \int_\Omega \left(1 + \|\nabla \widetilde{\mathbf{u}}_h\|^2\right)^{-\frac{3}{2}} \left[\left(1 + \|\nabla \widetilde{\mathbf{u}}_h\|^2\right) \|\nabla \widetilde{\mathbf{w}}_h\|^2 - \langle \nabla \widetilde{\mathbf{u}}_h, \nabla \widetilde{\mathbf{w}}_h \rangle^2\right] d\mathbf{x} \\
&\geq \int_\Omega \left(1 + \|\nabla \widetilde{\mathbf{u}}_h\|^2\right)^{-\frac{3}{2}} \|\nabla \widetilde{\mathbf{w}}_h\|^2 \, d\mathbf{x} \\
&> 0
\end{aligned}
$$

Aplicando el penúltimo teorema se tiene que $\widetilde{J}''$ es estrictamente convexa y por el último teorema se tiene que el problema (P) tiene solución única.

⊠

7. Métodos de optimización

7.1. Métodos de relajación y gradiente para problemas sin restricciones

Se comienza por generalizar la noción de funcional cuadrático en $\mathbb{R}^n$ con una matriz definida positiva. Esta extensión es muy adecuada para el estudio de los métodos considerados, lo cual conduce a pruebas de convergencia particularmente simples.
Un funcional $J : V \to \mathbb{R}$ definida en un espacio de Hilbert V se llama *elíptico* si es continuamente derivable en V y si existe una constante, que se convence a denotar α, tal que

$$\alpha > 0 \text{ y } \langle \nabla J(\mathbf{v}) - \nabla J(\mathbf{u}), \mathbf{v} - \mathbf{u} \rangle \geq \alpha \left\| \mathbf{v} - \mathbf{u} \right\|^2 \text{ para todo } \mathbf{u}, \mathbf{v} \in V$$

El siguiente resultado reúne varias propiedades de funcionales elípticas, que se utilizarán constantemente en lo posterior.

• ***Teorema 7:***

(i) Un funcional elíptico $J : V \to \mathbb{R}$ es estrictamente convexo y coercitivo y verifica la desigualdad

$$J(\mathbf{v}) - J(\mathbf{u}) \geq \langle \nabla J(\mathbf{u}), \mathbf{v} - \mathbf{u} \rangle + \frac{\alpha}{2} \left\| \mathbf{v} - \mathbf{u} \right\|^2 \text{ para todo } \mathbf{u}, \mathbf{v} \in V$$

(ii) Si U es un conjunto no vacío, convexo y cerrado del espacio de Hilbert V, y si J es una función elíptica. Ahora, el problema: Encontar $\mathbf{u}$ tal que

$$\mathbf{u} \in U \subseteq V \text{ y } J(\mathbf{u}) = \inf_{\mathbf{v} \in U} J(\mathbf{v}) \tag{P}$$

tiene una solución y sólo una.

(iii) Se supone que el conjunto U es convexo y el funcional J es elíptico. Entonces un elemento $\mathbf{u} \in U$ es la solución del problema P si y solo si verifica

$$\langle \nabla J(\mathbf{u}), \mathbf{v} - \mathbf{u} \rangle \geq 0 \text{ para todo } \mathbf{v} \in U$$

en el caso general, o si se tiene que

$$\nabla J(\mathbf{u}) = 0 \text{ si } U = V$$

(iv) Un funcional dos veces derivable en V es elíptico si y solo si

$$\left\langle \nabla^2 J(\mathbf{u})\mathbf{w}, \mathbf{w} \right\rangle \geq \alpha \left\| \mathbf{w} \right\|^2 \text{ para todo } \mathbf{w} \in V$$

○ *Demostración:*

Una funcional elíptico es por definición una vez continuamente derivable, la aplicación

de la fórmula de Taylor con resto permite escribir:

$$\begin{aligned}
J(\mathbf{v}) - J(\mathbf{u}) &= \int_0^1 \langle \nabla J(\mathbf{u} + t(\mathbf{v}-\mathbf{u})), \mathbf{v}-\mathbf{u}\rangle \, dt \\
&= \langle \nabla J(\mathbf{u}), \mathbf{v}-\mathbf{u}\rangle + \int_0^1 \langle \nabla J(\mathbf{u} + t(\mathbf{v}-\mathbf{u})) - \nabla J(\mathbf{u}), \mathbf{v}-\mathbf{u}\rangle \, dt \\
&\geq \langle \nabla J(\mathbf{u}), \mathbf{v}-\mathbf{u}\rangle + \int_0^1 \alpha t \left\|\mathbf{v}-\mathbf{u}\right\|^2 dt \\
&= \langle \nabla J(\mathbf{u}), \mathbf{v}-\mathbf{u}\rangle + \frac{\alpha}{2}\left\|\mathbf{v}-\mathbf{u}\right\|^2
\end{aligned}$$

De esta reducción se sigue, en primer lugar, que el funcional es estrictamente convexo ya que

$$J(\mathbf{v}) \geq J(\mathbf{u}) + \langle \nabla J(\mathbf{u}), \mathbf{v}-\mathbf{u}\rangle \text{ para todo } \mathbf{u},\mathbf{v} \in V, \mathbf{u} \neq \mathbf{v}$$

y, en segundo lugar, que el funcional es coercitivo, ya que

$$\begin{aligned}
J(\mathbf{v}) &\geq J(0) + \langle \nabla J(0), \mathbf{v}\rangle + \frac{\alpha}{2}\left\|\mathbf{v}\right\|^2 \\
&\geq J(0) - \left\|\nabla J(0)\right\| \left\|\mathbf{v}\right\| + \frac{\alpha}{2}\left\|\mathbf{v}\right\|^2
\end{aligned}$$

la existencia de una solución del problema (P) resulta del *teorema 5*, que se puede aplicar ya que el funcional es coercitivo; la unicidad resulta de su convexidad estricta. Si la función J es dos veces derivable en V y elíptica, se puede escribir

$$\begin{aligned}
\left\langle \nabla^2 J(\mathbf{u})\mathbf{w}, \mathbf{w}\right\rangle &= \lim_{\theta \to 0} \frac{\langle \nabla J(\mathbf{u}+\theta\mathbf{w}) - \nabla J(\mathbf{u}), \mathbf{w}\rangle}{\theta} \\
&= \lim_{\theta \to 0} \frac{\langle \nabla J(\mathbf{u}+\theta\mathbf{w}) - \nabla J(\mathbf{u}), \theta\mathbf{w}\rangle}{\theta^2} \\
&\geq \alpha \left\|\mathbf{w}\right\|^2
\end{aligned}$$

Recíprocamente, la fórmula de Taylor-Maclaurin aplicada a la función

$$f : \mathbf{w} \in V \longrightarrow f(\mathbf{w}) \stackrel{\text{def}}{=} \langle \nabla J(\mathbf{w}), \mathbf{v}-\mathbf{u}\rangle \in \mathbb{R}$$

Haciendo fijos los vectores $\mathbf{v}$ y $\mathbf{w}$, muestra que

$$\begin{aligned}
\langle \nabla J(\mathbf{v}) - \nabla J(\mathbf{u}), \mathbf{v}-\mathbf{u}\rangle &= f(\mathbf{v}) - f(\mathbf{u}) \\
&= f'\left(\mathbf{u} + \theta\left(\mathbf{v}-\mathbf{u}\right)\right)\left(\mathbf{v}-\mathbf{u}\right), 0 < \theta < 1 \\
&= \left\langle \nabla^2 J\left(\mathbf{u} + \theta\left(\mathbf{v}-\mathbf{u}\right)\right)\left(\mathbf{v}-\mathbf{u}\right), \mathbf{v}-\mathbf{u}\right\rangle \\
&\geq \alpha \left\|\mathbf{v}-\mathbf{u}\right\|^2
\end{aligned}$$

■

∘ *Observaciones:*

(i) En la última parte de la demostración, obviamente no se trata de escribir la fórmula de Taylor-Maclaurin para derivadas, ya que esta fórmula solo se aplica a funciones con valores en $\mathbb{R}$.

(ii) Un funcional cuadrático en $\mathbb{R}^n$:

$$J : \mathbf{v} \in \mathbb{R}^n \longrightarrow J(\mathbf{v}) = \frac{1}{2}\langle A\mathbf{v}, \mathbf{v}\rangle - \langle \mathbf{b}, \mathbf{v}\rangle, A = A^t$$

es elíptica si y solo si la matriz A es definida positiva. Se sigue que

$$\left\langle \nabla^2 J(\mathbf{u})\mathbf{w}, \mathbf{w}\right\rangle = \langle A\mathbf{w}, \mathbf{w}\rangle \geq \lambda_1 \|\mathbf{w}\|^2 \text{ para todo } \mathbf{u}, \mathbf{w} \in \mathbb{R}^n$$

donde λ_1 denota el valor propio más pequeño de la matriz A. Notar de pasada la desigualdad

$$\left\langle \nabla^2 J(\mathbf{u})\mathbf{w}, \mathbf{w}\right\rangle = \langle A\mathbf{w}, \mathbf{w}\rangle \leq \lambda_n \|\mathbf{w}\|^2 \text{ para todo } \mathbf{u}, \mathbf{w} \in \mathbb{R}^n$$

donde $\lambda_n = ||A||_2$, denota el mayor valor propio de la matriz A.

(iii) De manera similar, un funcional cuadrático en un espacio V de Hilbert,

$$J : \mathbf{v} \in V \longrightarrow J(\mathbf{v}) = \frac{1}{2}a(\mathbf{v}, \mathbf{v}) - f(\mathbf{v})$$

es elíptico si y solo si existe una constante α tal que

$$\alpha > 0 \text{ y } \left\langle \nabla^2 J(\mathbf{u})\mathbf{v}, \mathbf{v}\right\rangle = a(\mathbf{v}, \mathbf{v}) \geq \alpha \|\mathbf{v}\|^2 \text{ para todo } \mathbf{v} \in V$$

Es precisamente bajo este supuesto que se estableció el*teorema 6*

□

Se muestra ahora la descripción y luego el análisis de algunos algoritmos para resolver un problema de optimización *sin restricciones*: Dado un funcional J definido en un espacio vectorial V, encontrar $\mathbf{u}$ tal que

$$\mathbf{u} \in U \subseteq V \text{ y } J(\mathbf{u}) = \inf_{\mathbf{v}\in U} J(\mathbf{v}) \tag{P}$$

Se trata de *métodos iterativos* donde, partiendo de un vector inicial $\mathbf{u}_0$ arbitrario, se construye una sucesión de vectores $\mathbf{u}_k, k \geq 0$. Naturalmente el objetivo es la construcción de métodos *convergentes*, en el sentido de que, para cualquier vector inicial $\mathbf{u}_0$, la sucesión $(\mathbf{u}_k)_{k\geq 0}$ converge hacia una solución del problema (P).
Para construir el vector $\mathbf{u}_{k+1}$ una primera idea consiste en reducir a un problema "fácil de resolver numéricamente", a saber, un problema de minimización para una función de un valor real. Para eso, se propone lo siguiente:

(i) Darse una dirección "de bajada" en el punto $\mathbf{u}_k$, por medio de un vector $\mathbf{d}_k$ distinto de cero

(ii) Encontrar el mínimo de la restricción del funcional J a la derecha que pasa por el punto $\mathbf{u}_k$ y paralelo al vector $\mathbf{d}_k$: esto define el vector $\mathbf{u}_{k+1}$ sólo si la variable: Hallar $\varrho(\mathbf{u}_k, \mathbf{d}_k)$ tal que

$$\varrho(\mathbf{u}_k, \mathbf{d}_k) \in \mathbb{R}$$

y

$$J(\mathbf{u}_k + \varrho(\mathbf{u}_k, \mathbf{d}_k)\mathbf{d}_k) = \inf_{\varrho\in\mathbb{R}} J(\mathbf{u}_k + \varrho\mathbf{d}_k)$$

Estas consideraciones se ilustran en el caso de la dimensión dos en la *figura 5*. La superficie que representa un funcional elíptico tiene la apariencia de un paraboloide cuyas posiciones horizontales tienen la forma de elipses, lo que explica además el término "funcional elíptico".

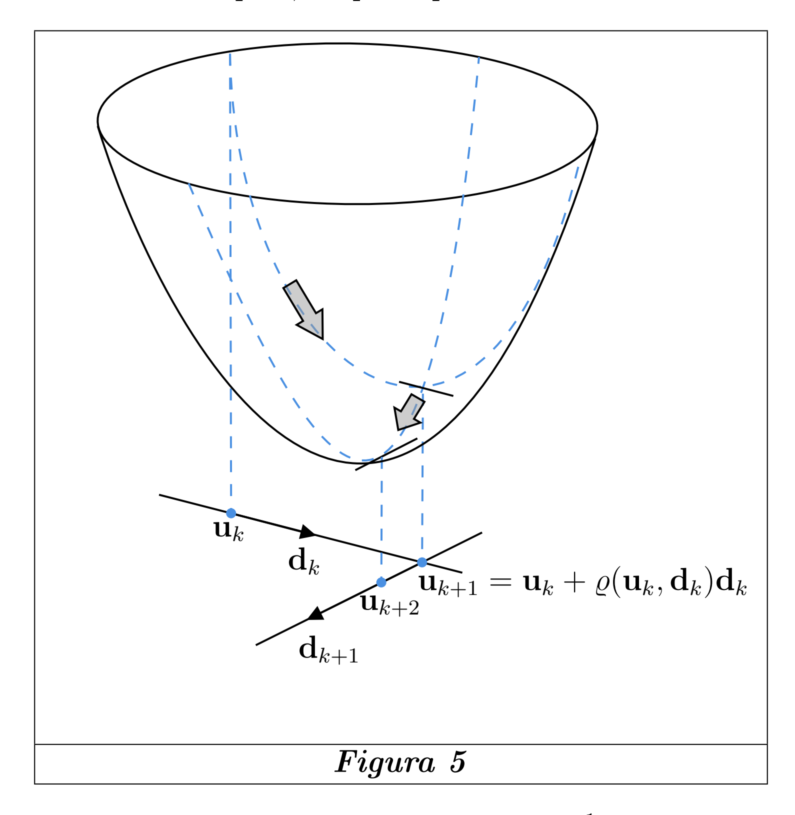


Figura 5

En el caso de una funcional cuadrático elíptico $J(\mathbf{v}) = \frac{1}{2}a(\mathbf{v},\mathbf{v}) - f(\mathbf{v})$ es esencial tener en cuenta que la determinación del punto $\mathbf{u}_{k+1}$ está inmediatamente ligada al vector $\mathbf{d}_k$, ya que la función

$$\varrho \in \mathbb{R} \longrightarrow J(\mathbf{u}_k + \varrho \mathbf{d}_k) = \frac{\varrho^2}{2} a(\mathbf{d}_k, \mathbf{d}_k) + \varrho(\nabla J(\mathbf{u}_k), \mathbf{d}_k) + J(\mathbf{u}_k)$$

es un trinomio cuadrático (el coeficiente $a(\mathbf{d}_k, \mathbf{d}_k)$ es > 0).
En el caso donde $V = \mathbb{R}^n$, la forma más sencilla de definir las direcciones sucesivas de descenso consiste en imponer de antemano, siendo una elección "canónica" a este respecto, naturalmente, la de las direcciones de los ejes de coordenadas, tomadas de manera "cíclica"; esta es la idea del *método de relajación*: A partir de un vector inicial $\mathbf{u}_0$, cada vector $\mathbf{u}_{k+1} = \left(u_i^{k+1}\right)_{i=1}^n$ construido (cuando es posible, naturalmente) a partir del vector $\mathbf{u}_k = \left(u_i^k\right)_{i=1}^n$ calculando sucesivamente sus componentes resolviendo los siguientes problemas de minimización a una variable (cada "nueva" componente calculada se ha encerrado entre corchetes):

$$\left\{\begin{array}{lcl}
J(\left[u_1^{k+1}\right], u_2^k, u_3^k, \quad \ldots \quad, u_{n-1}^k, u_n^k) & = & \operatorname{ínf}_{\zeta\in\mathbb{R}} J\left(\zeta, u_2^k, u_3^k, \ldots, u_n^k\right) \\
J(u_1^{k+1}, \left[u_2^{k+1}\right], u_3^k, \quad \ldots \quad, u_{n-1}^k, u_n^k) & = & \operatorname{ínf}_{\zeta\in\mathbb{R}} J\left(u_1^{k+1}, \zeta, u_3^k, \ldots, u_n^k\right) \\
 & \vdots & \\
J(u_1^{k+1}, u_2^{k+1}, u_3^{k+1}, \quad \ldots \quad, u_{n-1}^{k+1}, \left[u_n^{k+1}\right]) & = & \operatorname{ínf}_{\zeta\in\mathbb{R}} J\left(u_1^{k+1}, u_2^{k+1}, u_3^{k+1}, \ldots, u_{n-1}^{k+1}, \zeta\right)
\end{array}\right.$$

Es conveniente, con vistas a la demostración siguiente, introducir los vectores "intermedios"

$\mathbf{u}_{k;l}, 0 \leq l \leq n$, definidos por

$$\begin{array}{rcl} \mathbf{u}_k = \mathbf{u}_{k;0} & = & \left(u_1^k, \dots, u_n^k\right), \\ \mathbf{u}_{k;1} & = & \left(u_1^{k+1}, u_2^k \dots, u_n^k\right) \\ & \vdots & \\ \mathbf{u}_{k;l} & = & \left(u_1^{k+1}, \dots, u_l^{k+1}, u_{l+1}^k, \dots, u_n^k\right) \\ & \vdots & \\ \mathbf{u}_{k;n} & = & \left(u_1^{k+1}, \dots, u_n^{k+1}\right) = \mathbf{u}_{k+1} \end{array}$$

para que los problemas de minimización anteriores queden en la forma equivalente:

$$\begin{array}{rcl} J(\mathbf{u}_{k;1}) & = & \text{ínf}_{\varrho \in \mathbb{R}} J(\mathbf{u}_{k;0} + \varrho \mathbf{e}_1) \\ & \vdots & \\ J(\mathbf{u}_{k;l}) & = & \text{ínf}_{\varrho \in \mathbb{R}} J(\mathbf{u}_{k;l-1} + \varrho \mathbf{e}_l) \\ & \vdots & \\ J(\mathbf{u}_{k;n}) & = & \text{ínf}_{\varrho \in \mathbb{R}} J(\mathbf{u}_{k;n-1} + \varrho \mathbf{e}_n) \end{array}$$

donde $(\mathbf{e}_l)$ denota la base canónica de $\mathbb{R}^n$. Se sujeta a la derivabilidad del funcional J, se deducen de ella las condiciones necesarias, y suficientes si además es convexa, de mínimo

$$\partial_l J(\mathbf{u}_{k;l}) = 0, 1 \leq l \leq n$$

usando la notación para primeras derivadas parciales

$$\partial_l J(\mathbf{v}) = J'(\mathbf{v})\mathbf{e}_l = \langle \nabla J(\mathbf{v}), \mathbf{e}_l \rangle, 1 \leq l \leq n$$

• ***Teorema 8:*** Si el funcional $J : \mathbb{R}^n \longrightarrow \mathbb{R}$ es elíptico, el método de relajación converge.

○ *Demostración:*

(i) Cada función

$$\varphi_{k;l} : \varrho \in \mathbb{R} \longrightarrow \varphi_{k;l}(\varrho) \stackrel{\text{def}}{=} J(\mathbf{u}_{k;l-1} + \varrho \mathbf{e}_l)$$

siendo coercitivo y estrictamente convexo, admite en mínimo y sólo uno. Cada sucesión $(\mathbf{u}_{k;l})_{k \geq 0}$,$1 \leq l \leq n$, por lo tanto, está bien definida, en particular, la sucesión $(\mathbf{u}_k)_{k \geq 0}$. Se escribe

$$J(\mathbf{u}_k) - J(\mathbf{u}_{k+1}) = J(\mathbf{u}_{k;0}) - J(\mathbf{u}_{k;n}) = \sum_{l=1}^{n} (J(\mathbf{u}_{k;l-1}) - J(\mathbf{u}_{k;l}))$$

y, según la hipótesis de la eliplicidad (teorema 8.4-1):

$$J(\mathbf{u}_{k;l-1}) - J(\mathbf{u}_{k;l}) \geq \langle \nabla J(\mathbf{u}_{k;l}), \mathbf{u}_{k;l-1} - \mathbf{u}_{k;l} \rangle + \frac{\alpha}{2} \parallel \mathbf{u}_{k;l-1} - \mathbf{u}_{k;l} \parallel^2$$

como

$$\langle \nabla J(\mathbf{u}_{k;l}), \mathbf{u}_{k;l-1} - \mathbf{u}_{k;l} \rangle = \partial_l J(\mathbf{u}_{k;l}) \left(u_l^k - u_l^{k+1}\right) = 0, 1 \leq l \leq n$$

y como

$$\parallel \mathbf{u}_{k;l-1} - \mathbf{u}_{k;l} \parallel^2 = \left|u_l^k - u_l^{k+1}\right|^2, 1 \leq l \leq n$$

se obtiene finalmente

$$J(\mathbf{u}_k) - J(\mathbf{u}_{k+1}) \geq \frac{\alpha}{2} \sum_{l=1}^{n} \left|u_l^k - u_l^{k+1}\right|^2 = \frac{\alpha}{2} \parallel \mathbf{u}_k - \mathbf{u}_{k+1} \parallel^2$$

(ii) Como la sucesión $(J(\mathbf{u}_k))_{k\geq 0}$ es decreciente y acotada inferiormente, se deduce de (i):

$$\lim_{k\to\infty} \| \mathbf{u}_k - \mathbf{u}_{k+1} \| = 0$$

y por lo tanto, a priori

$$\lim_{k\to\infty} \| \mathbf{u}_{k;l} - \mathbf{u}_{k+1} \| = 0, 1 \leq l \leq n-1$$

(iii) Usando la elipticidad del funcional y la caracterización $\nabla J(\mathbf{u}) = 0$ del mínimo $\mathbf{u}$ (*teorema 7*), se obtiene

$$\begin{array}{rcl} \alpha \| \mathbf{u}_{k+1} - \mathbf{u} \|^2 & \leq & \langle \nabla J(\mathbf{u}_{k+1}) - \nabla J(\mathbf{u}), \mathbf{u}_{k+1} - \mathbf{u} \rangle \\ & = & \langle \nabla J(\mathbf{u}_{k+1}), \mathbf{u}_{k+1} - \mathbf{u} \rangle \\ & = & \sum_{l=1}^{n} \partial_l J(\mathbf{u}_{k+1}) \left(u_l^{k+1} - u_l \right) \end{array}$$

de lo que se deduce, con las caracterizaciones $\partial_l J(\mathbf{u}_{k;l})$:

$$\| \mathbf{u}_{k+1} - \mathbf{u} \| \leq \frac{1}{\alpha} \sum_{l=1}^{n} |\partial_l J(\mathbf{u}_{k+1})| = \frac{1}{\alpha} \sum_{l=1}^{n} |\partial_l J(\mathbf{u}_{k+1}) - \partial_l J(\mathbf{u}_{k;l})|$$

(iv) Como cada sucesión $(J(\mathbf{u}_{k;l}))_{k\geq 0}$ decreciente por construcción, cada sucesión $(\mathbf{u}_{k;l})_{k\geq 0}, 1 \leq l \leq n$, está acotada ya que el funcional es coercitivo. Como además cada primera derivada parcial $\partial_l J$ es uniformemente continua en los compactos de $\mathbb{R}^n$

$$\lim_{k\to\infty} \| \mathbf{u}_{k;l} - \mathbf{u}_{k+1} \| = 0 \Rightarrow \lim_{k\to\infty} |\partial_l J(\mathbf{u}_{k;l}) - \partial_l J(\mathbf{u}_{k+1})| = 0, 1 \leq l \leq n$$

y la convergencia se sigue de (iii)

■

- *Observaciones:*

(i) La derivabilidad del funcional es un supuesto esencial, a continuación, se considera el ejemplo del funcional

$$J : \mathbf{v} = (v_1, v_2) \in \mathbb{R}^2 \longrightarrow J(v_1, v_2) = v_1^2 + v_2^2 - 2(v_1 + v_2) + 2|v_1 - v_2|$$

que es coercitivo, estrictamente convexo, "casi-cuadrático" pero no diferenciable: con la elección, $\mathbf{u}_0 = (0,0)$ para el vector inicial, el método de relajación conduce a la sucesión estacionaria $(0,0) = \mathbf{u}_0 = \mathbf{u}_1 = \ldots = \mathbf{u}_k = \ldots$, mientras $\inf_{\mathbf{v}\in\mathbb{R}^2} J(\mathbf{v}) = J(1,1)$. No obstante, se puede establecer la convergencia para funcionales no derivables del tipo

$$J(\mathbf{v}) = J_0(\mathbf{v}) + \sum_{i=1}^{n} \alpha_i |v_i|, \alpha_i \geq 0$$

siendo la función J elíptica.

(ii) Se puede probar la analogía del *teorema 8* bajo las siguientes hipótesis más generales (pero es un poco más delicada): el funcional es continuamente derivable, estrictamente convexo y coercitivo.

(*iii*) Esta es la hipótesis de la *dimensión finita*, que, por intermedio de la *continuidad uniforme*, juega un papel esencial en la demostración. De hecho, sin esta última propiedad, las últimas implicaciones de la demostración ya no son necesariamente verdaderas.

(*iv*) La estimación obtenida en (*iii*) proporciona un incremento a priori del error $\| \mathbf{u}_k - \mathbf{u} \|$, en principio enteramente deducible de los datos.

□

Se considera el caso particular de un funcional cuadrático

$$J(\mathbf{v}) = \frac{1}{2}\langle A\mathbf{v}, \mathbf{v}\rangle - \langle \mathbf{b}, \mathbf{v}\rangle = \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}v_i v_j - \sum_{i=1}^{n} b_i v_i$$

Se puede aplicar el teorema 8.4-2 si la matriz simétrica $A = (a_{ij})$ es definida positiva. Ya que

$$\partial_l J(\mathbf{v}) = \sum_{j=1}^{n} a_{lj}v_j - b_l, 1 \leq l \leq n$$

se deduce (con las notaciones usadas arriba)

$$\begin{array}{rcll} \partial_1 J(\mathbf{u}_{k;1}) & = & a_{11}\left[u_1^{k+1}\right] + a_{12}u_2^k + \cdots + a_{1n}u_n^k - b_1 & = 0 \\ \partial_1 J(\mathbf{u}_{k;2}) & = & a_{21}u_1^{k+1} + a_{22}\left[u_2^{k+1}\right] + a_{23}u_3^k + \cdots + a_{2n}u_n^k - b_2 & = 0 \\ & \vdots & & \\ \partial_1 J(\mathbf{u}_{k;n}) & = & a_{n1}u_1^{k+1} + a_{n2}u_2^{k+1} + \cdots + a_{nn-1}u_{n-1}^{k+1} + a_{nn}\left[u_n^{k+1}\right] - b_n & = 0 \end{array}$$

Se nota que es constante encontrar exactamente el *método de Gauss-Seidel para la solución del sistema lineal* $A\mathbf{u} = \mathbf{b}$; el *teorema 8* proporciona así una nueva prueba de la convergencia de este método cuando la matriz A es simétrica definida positiva. Siendo el método de Gauss-Seidel un caso particular del *método de relajación para la resolución de los sistemas lineales*, la terminología empleada está, pues, parcialmente justificada.
Considerando de nuevo el problema general de la optimización soles restringida en el caso donde $V = \mathbb{R}^n$: Encuentre $\mathbf{u} \in \mathbb{R}^n$ tal que $J(\mathbf{u}) = \text{ínf}_{\mathbf{v}\in\mathbb{R}^n} J(\mathbf{v})$. Parece intuitivamente claro que la convergencia de un método iterativo debería ser mucho mejor ya que las diferencias $J(\mathbf{u}_k) - J(\mathbf{u_{k+1}})$ son grandes, y en este sentido, la elección impuesta de las direcciones de los ejes de coordenadas ciertamente no es óptima.
Para hacer la diferencia $J(\mathbf{u}_k) - J(\mathbf{u_{k+1}})$ lo más grande posible, la idea más inmediata consiste en elegir como dirección de descenso la de mayor descenso *local*, es decir, el opuesto al gradiente $\nabla J(\mathbf{u}_k)$. Se recuerda de paso la justificación de esta última afirmación: Por definición del gradiente, puede escribirse

$$J(\mathbf{u}_k + \mathbf{w}) = J(\mathbf{u}_k) + \langle \nabla J(\mathbf{u}_k), \mathbf{w}\rangle + \| \mathbf{w} \| \varepsilon(\mathbf{w}), \lim_{\mathbf{w}\to 0} \varepsilon(\mathbf{w}) = 0$$

de modo que, si $\nabla J(\mathbf{u}_k) \neq 0$, la parte principal del incremento de la función J se incrementa en módulo por el producto $\| \nabla J(\mathbf{u}_k) \| \| \mathbf{w} \|$ (desigualdad de Cauchy-Schwarz), con igualdad si y sólo si los dos vectores $\nabla J(\mathbf{w})$ y $\mathbf{w}$ son proporcionales.
Se disponen pues de todos los elementos necesarios para la definición del método correspondiente a esta elección de dirección de descenso, denominado *método del gradiente con paso óptimo*:

A partir de un vector inicial $\mathbf{u}_0$ cada vector $\mathbf{u}_{k+1}$ se construye (cuando es posible, naturalmente) a partir del vector $\mathbf{u}_k, k \geq 0$, por las relaciones

$$\begin{cases} J(\mathbf{u}_k - \varrho(\mathbf{u}_k)\nabla J(\mathbf{u}_k)) & = \inf_{\varrho \in \mathbb{R}} J(\mathbf{u}_k - \varrho \nabla J(\mathbf{u}_k)) \\ \mathbf{u}_k & = \mathbf{u}_k - \varrho(\mathbf{u}_k)\nabla J(\mathbf{u}_k) \end{cases}$$

El signo "menos" delante de la variable ϱ recuerda que la dirección de descenso es en dirección opuesta a la del gradiente; se espera un valor > 0 para el número $\varrho(\mathbf{u}_k)$

○ *Observación:*

Contrariamente a la intuición, la dirección $\mathbf{d}_k = -\nabla J(\mathbf{u}_k)$ no es necesariamente óptima: ¡la siguiente sección es muy instructiva a este respecto!.

□

Antes de pasar al estudio de la convergencia del método del gradiente con el paso óptimo, se considera una definición general: cualquier método iterativo para el cual el punto $\mathbf{u}_{k+1}$ es de la forma

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \varrho_k \nabla J(\mathbf{u}_k), \varrho_k > 0$$

se denomina un *método de gradiente*. El método anterior es por lo tanto un primer caso particular; otros dos se estudian a continuación.

• ***Teorema 9:*** Se supone $V \longrightarrow \mathbb{R}^n$ y un funcional elíptico. Entonces el método del gradiente con paso óptimo converge.

○ *Demostración:*

(i) Sin pérdida de generalidad, se supone que $\nabla J(\mathbf{u}_k) \neq 0$ para todo $k \geq 0$; de lo contrario, el método es convergente en un número finito de iteraciones. Cada función

$$\varphi_k : \varrho \in \mathbb{R} \longrightarrow \varphi_k(\varrho) \stackrel{\text{def}}{=} J(\mathbf{u}_k - \varrho \nabla J(\mathbf{u}_k))$$

siendo coercitiva y estrictamente convexa, admite un mínimo y sólo uno, caracterizado por la relación $\varphi'_k(\varrho(\mathbf{u}_k)) = 0$. Como

$$\varphi'_k(\varrho) = -(\nabla J(\mathbf{u}_k - \varrho \nabla J(\mathbf{u}_k)), \nabla J(\mathbf{u}_k))$$

se deduce la relación

$$\langle \nabla J(\mathbf{u}_{k+1}), \nabla J(\mathbf{u}_k) \rangle = 0$$

que muestra que *dos direcciones sucesivas de descenso son ortogonales*. Como $\mathbf{u}_k = \mathbf{u}_k - \varrho(\mathbf{u}_k)\nabla J(\mathbf{u}_k)$, también se tiene

$$\langle \nabla J(\mathbf{u}_{k+1}), \mathbf{u}_{k+1} - \mathbf{u}_k \rangle = 0$$

por lo tanto, por aplicación de la primera desigualdad del *teorema 7*,

$$J(\mathbf{u}_k) - J(\mathbf{u}_{k+1}) \geq \frac{\alpha}{2} \parallel \mathbf{u}_k - \mathbf{u}_{k+1} \parallel^2$$

(ii) Como la sucesión $J(\mathbf{u}_k)_{k \geq 0}$ es decreciente (por construcción) y acotada inferiormente (por $J(\mathbf{u})$), se deduce

$$\lim_{k \to \infty} (J(\mathbf{u}_k) - J(\mathbf{u}_{k+1})) = 0$$

una relación que, unida a la desigualdad anterior, muestra que

$$\lim_{k \to \infty} \parallel \mathbf{u}_k - \mathbf{u}_{k+1} \parallel = 0$$

(*iv*) Como la sucesión $J(\mathbf{u}_k)_{k\geq 0}$ es decreciente, la sucesión $(\mathbf{u}_k)_{k\geq 0}$ está acotada ya que el funcional es coercitivo (*teorema 7*). La función derivada J', continua por hipótesis, es por lo tanto *uniformemente continua en compactos*. Se sigue entonces de (*ii*) que

$$\lim_{k\to\infty} \parallel \nabla J(\mathbf{u}_k) - \nabla J(\mathbf{u}_{k+1}) \parallel = 0$$

y por tanto, según (*iii*) que

$$\lim_{k\to\infty} \nabla J(\mathbf{u}_k) = 0$$

(*v*) Se prueba finalmente la convergencia. Se escribe

$$\alpha \parallel \mathbf{u}_k - \mathbf{u} \parallel^2 \leq \langle \nabla J(\mathbf{u}_k) - \nabla J(\mathbf{u}), \mathbf{u}_k - \mathbf{u}\rangle = \langle \nabla J(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\rangle \leq \parallel \nabla J(\mathbf{u}_k) \parallel \parallel \mathbf{u}_k - \mathbf{u} \parallel$$

utilizando sucesivamente la hipótesis de elipticidad del funcional, se tiene entonces la relación $\nabla J(\mathbf{u}) = 0$. De esta manera, se obtiene

$$\parallel \mathbf{u}_k - \mathbf{u} \parallel \leq \frac{1}{\alpha} \parallel \nabla J(\mathbf{u}_k) \parallel$$

y se sigue la conclusión de la propiedad establecida en (*iv*).

■

∘ *Observaciones:*

(*i*) Igual que para el método de relajación. La hipotensión de dimensión finita jugó un papel esencial en esta demostración.

(*ii*) Se puede probar el análogo del *teorema 8* bajo las siguientes hipótesis más generales: el funcional es una vez continuamente derivable, estrictamente convexo y coercitivo.

(*iii*) Se puede dar otra demostración de convergencia, que probablemente se aplique a situaciones más genéricas: la sucesión $(\mathbf{u}_k)$ está acotada, sea $(\mathbf{u}_{k'})$ una sucesión extraída que converge en un elemento $\mathbf{u}'$. De la continuidad de la aplicación derivada, se deduce

$$\nabla J(\mathbf{u}') = \lim_{k'\to\infty} \nabla J(\mathbf{u}'_{k'}) = 0$$

según (*iv*). Como la solución del problema está caracterizada por la relación $\nabla J(\mathbf{u}) = 0$, se deduce $\mathbf{u} = \mathbf{u}'$ por un lado, y la convergencia de toda la sucesión $(\mathbf{u}_k)$ por otro lado, siendo el límite único.

(*iv*) Si la demostración de la convergencia dada en la parte (*v*) es particular para las funciones elípticas, tiene la ventaja de proporcionar una *mayorización para el error* $\parallel \mathbf{u}_k - \mathbf{u} \parallel$, en principio totalmente calculable a priori.

□

En el caso de un *funcional cuadrático elíptico*:

$$J(\mathbf{v}) = \frac{1}{2}\langle A\mathbf{v}, \mathbf{v}\rangle - \langle \mathbf{b}, \mathbf{v}\rangle$$

la ortogonalidad de los vectores $\nabla J(\mathbf{u}_k)$ y $\nabla J(\mathbf{u}_{k+1})$ puede ser una ventaja para calcular el número $\varrho(\mathbf{u}_k)$. Sabiendo que $\nabla J(\mathbf{u}) = A\mathbf{v} - \mathbf{b}$, se escribe

$$0 = \langle \nabla J(\mathbf{u}_{k+1}), \nabla J(\mathbf{u}_k)\rangle = \langle A(\mathbf{u}_k - \varrho(\mathbf{u}_k)(A\mathbf{u}_k - \mathbf{b})) - \mathbf{b}, A\mathbf{u}_k - \mathbf{b}\rangle$$

de donde se deduce

$$\varrho(\mathbf{u}_k) = \frac{\| \mathbf{w}_k \|^2}{\langle A\mathbf{w}_k, \mathbf{w}_k \rangle} \text{ ,donde } \mathbf{w}_l \stackrel{\text{def}}{=} A\mathbf{u}_k - \mathbf{b} = \nabla J(\mathbf{u}_k)$$

Una iteración del método queda entonces de la siguiente forma:

$$\begin{cases} \text{Cálculo de vectores} & \mathbf{w}_k = A\mathbf{u}_k - \mathbf{b} \\ \text{Cálculo de números} & \varrho(\mathbf{u}_k) = \frac{\|\mathbf{w}_k\|^2}{\langle A\mathbf{w}_k, \mathbf{w}_k\rangle} \\ \text{Cálculo de vectores} & \mathbf{u}_{k+1} = \mathbf{u}_k - \varrho(\mathbf{u}_k)\mathbf{w}_k \end{cases}$$

Se notará de paso que se trata de un nuevo *método iterativo de resolución de un sistema lineal* $A\mathbf{u} = \mathbf{b}$ cuya matriz A es simétrica y definida positiva. Tal método puede resultar interesante cuando el cálculo de un vector $A\mathbf{w}$, donde $\mathbf{w}$ es un vector conocido, es fácil. Este es esencialmente el caso de las *matrices dispersas (o huecas)*, especialmente aquellas que se obtienen durante la discretización de problemas de contorno. Se considera este punto con más detalle en el siguiente párrafo, sobre el método del gradiente conjugado.
Los métodos de relajación y de gradiente con paso óptimo tienen en común la búsqueda de mínimos de funciones de una variable. En particular para superar esta obligación se define el *método de gradiente de paso fijo*: a partir de un vector inicial $\mathbf{u}_0$ arbitrario, la sucesión $(\mathbf{u}_k)$ se define por

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \varrho\nabla J(\mathbf{u}_k), k \geq 0$$

el parámetro real ϱ que se determina "en el mejor de los casos". De manera más general,se define el *método de gradiente de con paso variable*, estableciendo

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \varrho_k\nabla J(\mathbf{u}_k), k \geq 0$$

, los parámetros reales ϱ_k se ajustan, por ejemplo, durante las iteraciones de acuerdo con criterios particulares. Tener en cuenta que el *método de gradiente de paso fijo* es un caso particular del *método del gradiente de paso variable.*
Se dan ahora condiciones suficientes de *convergencia* para funcionales elípticos. Su naturaleza también es fácil de predecir: el parámetro ϱ, y los parámetros ϱ_k deben estar en un intervalo compacto de la forma $[a, b], a > 0$. En otras palabras, "se desciende" efectivamente ($\varrho_k \geq a$) y no "demasiado lejos" ($\varrho_k \leq b$): Esto es lo que se trata de sugerir en la *figura 6.*

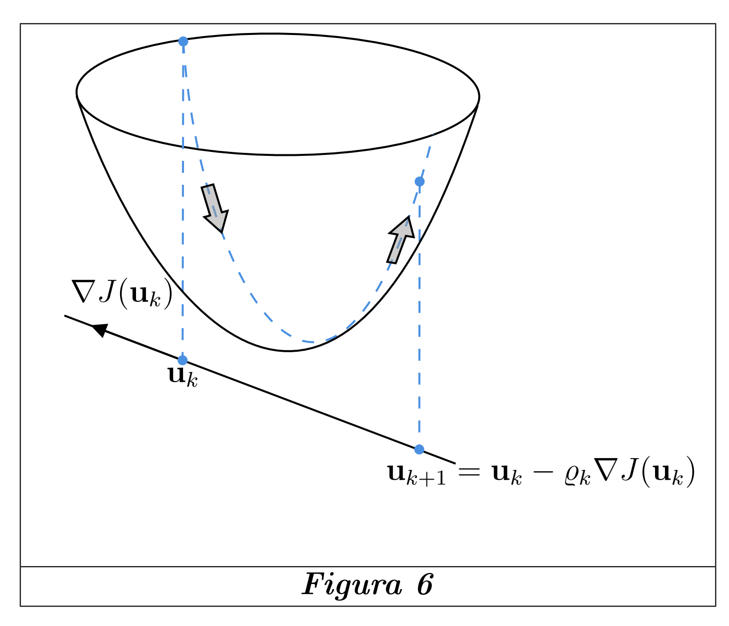


Figura 6

• ***Teorema 10:*** Sea V un espacio de Hilbert y $J : V \longrightarrow \mathbb{R}$ un funcional derivable en V. Se supone que hay dos constantes α y M tales que:

$$\begin{array}{ll} \alpha > 0 \quad \text{y} & \langle \nabla J(\mathbf{v}) - \nabla J(\mathbf{u}), \mathbf{v} - \mathbf{u} \rangle \geq \alpha \parallel \mathbf{v} - \mathbf{u} \parallel^2 \ \text{, para todo } \mathbf{u}, \mathbf{v} \in V \\ & \parallel \nabla J(\mathbf{v}) - \nabla J(\mathbf{u}) \parallel \leq M \parallel \mathbf{v} - \mathbf{u} \parallel \ \text{, para todo } \mathbf{u}, \mathbf{v} \in V \end{array}$$

Si existen dos números a y b tales que

$$0 < a \leq \varrho_k \leq b < \frac{2\alpha}{M^2} \text{ ,para todo entero } k \geq 0$$

el método de gradiente de paso variable converge, y la convergencia es geométrica: existe una constante $\beta = \beta(\alpha, M, a, b)$ tal que

$$\beta < 1 \text{ y } \parallel \mathbf{u}_k - \mathbf{u} \parallel \leq \beta^k \parallel \mathbf{u}_0 - \mathbf{u} \parallel$$

○ *Demostración*

Utilizando la caracterización $\nabla J(\mathbf{u}) = 0$ de mínimo, se puede escribir

$$\begin{array}{rcl} \parallel \mathbf{u}_{k+1} - \mathbf{u} \parallel^2 & = & \parallel \mathbf{u}_k - \mathbf{u} \parallel^2 - 2\varrho_k \langle \nabla J(\mathbf{u}_k) - \nabla J(\mathbf{u}), \mathbf{u}_k - \mathbf{u} \rangle + \varrho_k^2 \parallel \nabla J(\mathbf{u}_k) - \nabla J(\mathbf{u}) \parallel^2 \\ & \leq & \{1 - 2\alpha\varrho_k + M^2\varrho_k^2\} \parallel \mathbf{u}_k - \mathbf{u} \parallel^2 \end{array}$$

suponiendo que $\varrho_k > 0$. Dado el trinomio $\tau(\varrho) = 1 - 2\alpha\varrho + M^2\varrho^2$, es claro que

$$0 < a \leq \varrho_k \leq b < \frac{2\alpha}{M^2} \Rightarrow \left(1 - 2\alpha\varrho_k + M^2\varrho_k^2\right)^{\frac{1}{2}} \leq \beta \overset{\text{def}}{=} (\text{máx}\{\tau(a), \tau(b)\})^{\frac{1}{2}} < 1$$

como entonces

$$\parallel \mathbf{u}_{k+1} - \mathbf{u} \parallel \leq \beta \parallel \mathbf{u}_k - \mathbf{u} \parallel \leq \beta^{k+1} \parallel \mathbf{u}_0 - \mathbf{u} \parallel$$

se demuestra la convergencia geométrica

■

Luego una iteración del método se presenta de la siguiente forma:

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \varrho_k(A\mathbf{u}_k - \mathbf{b}), k \geq 0$$

y resulta del teorema de que el método es convergente si $0 < a \leq \varrho_k \leq b \leq 2\frac{\lambda_1}{\lambda_n^2}$, sean λ_1 y λ_n el menor y el mayor de los autovalores de la matriz definida positiva simétrica A. Se puede mejorar este resultado: de hecho, de la igualdad

$$\mathbf{u}_{k+1} - \mathbf{u} = (\mathbf{u}_k - \mathbf{u}) - \varrho_k A(\mathbf{u}_k - \mathbf{u}) = (I - \varrho_k A)(\mathbf{u}_k - \mathbf{u})$$

se deduce la mayorización

$$\parallel \mathbf{u}_{k+1} - \mathbf{u} \parallel \leq \parallel I - \varrho_k A \parallel_2 \parallel \mathbf{u}_k - \mathbf{u} \parallel$$

Siendo la matriz $(I - \varrho_k A)$ simétrica, su norma $\parallel \cdot \parallel_2$ tiene por expresión:

$$\parallel I - \varrho_k A \parallel_2 = \text{máx}\{|1 - \varrho_k\lambda_1|, |1 - \varrho_k\lambda_n|\}$$

La forma de la función (*figura 7*)

$$\mu : \varrho \in \mathbb{R} \longrightarrow \mu(\varrho) = \text{máx}\{|1 - \varrho\lambda_1|, |1 - \varrho\lambda_n|\}$$

muestra que

$$0 < a \leq \varrho_k \leq \tilde{b} < \frac{2}{\lambda_n} \Rightarrow \tilde{\beta} \overset{\text{def}}{=} \text{máx}\left\{\mu(a), \mu\left(\tilde{b}\right)\right\} < 1$$

y por lo tanto que

$$\| \mathbf{u}_{k+1} - \mathbf{u} \| \leq \tilde{\beta} \| \mathbf{u}_k - \mathbf{u} \| \leq \tilde{\beta}^{k+1} \| \mathbf{u}_0 - \mathbf{u} \|$$

Ahora es claro que la cota superior $\frac{2\lambda_1}{\lambda_n^2}$ indicada por el teorema es en general "mucho" más perceptible que la cota $\frac{2}{\lambda_n}$ ya que su relación es el de los autovalores extremos de la matriz A. Finalmente, se señalará que los valores "óptimos" del parámetro ϱ encontrados por los dos métodos para el método de gradiente de paso fijo son respectivamente $\frac{\lambda_1}{\lambda_n^2}$ y $\frac{2}{\lambda_1+\lambda_n}$ (*figura 7*)

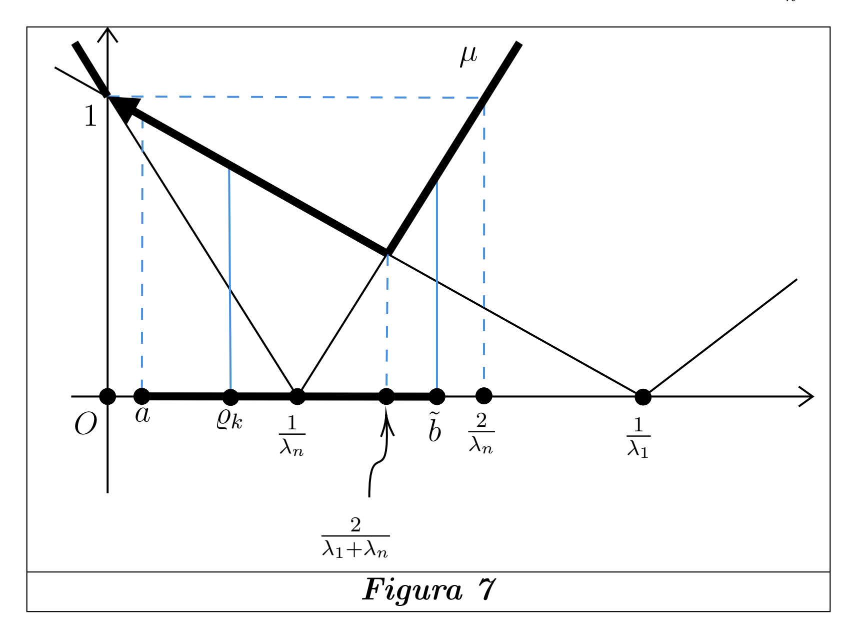


Figura 7

∘ *Observación*

La mejora anterior puede extenderse a funciones que no son necesariamente cuadráticas, por lo cual se puede establecer la convergencia tan pronto como $\varrho_k \in [a, \widetilde{b}]$, con $\widetilde{b} < \frac{2}{M}$, pero sin poder establecer su carácter geométrico.

□

Desde el punto de vista "numérico", el inconveniente de los métodos de gradiente es el cálculo del vector $\nabla J(\mathbf{u}_k)$ en cada iteración que, recordar, se utiliza para determinar la "siguiente" dirección de descenso, mientras que el inconveniente de los métodos de relajación y de gradiente con paso óptimo radican en la resolución de problemas de minimización univariados. Esta es la razón por la cual la elección real de un método depende en gran medida de la importancia relativa de estos aspectos "numéricos" y la velocidad esperada de convergencia.

7.2. Métodos de gradiente conjugado para problemas sin restricciones

Considerar el problema de la optimización sin restricciones: Encontrar $\mathbf{u}$ tal que

$$\mathbf{u} \in U \subseteq V \text{ y } J(\mathbf{u}) = \inf_{\mathbf{v} \in U} J(\mathbf{v}) \tag{P}$$

Como métodos de aproximación basados en la minimización de funciones a una variable en direcciones de descenso apropiadas, se ha estudiado el método de relajación y el método del

gradiente de paso óptimo. Las primeras direcciones sucesivas de descenso utilizadas fueron direcciones fueron impuestas de antemano (las de los ejes de coordenadas), independientemente del funcional J. El segundo utiliza en cada iteración una dirección "localmente óptima" (la del gradiente), esta vez relacionada con el funcional considerado; por lo tanto, se puede esperar una convergencia más rápida. Con el fin de mejorar aún más la convergencia, es evidente que hay que hacer esfuerzos para utilizar más información sobre el funcional para definir la dirección de los vectores $\mathbf{u}_{k+1} - \mathbf{u}_k$. Este es el caso, por ejemplo, del método de Newton, que se presenta como:

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \left(\nabla^2 J(\mathbf{u}_k)\right)^{-1} \nabla J(\mathbf{u}_k), k \geq 0$$

Si este método no requiere la solución de problemas de minimización de una sola variable, su principal inconveniente es la resolución de los sistemas lineales de matriz $\nabla^2 J(\mathbf{u}_k)$ para cada iteración, que es muy ajustable numéricamente. Sin embargo, es posible encontrar direcciones de descenso mejoradas en comparación con la pendiente sin recurrir a a las segundas derivadas del funcional.

Para convencerse de esto, se considera el caso, muy simple pero muy informativo, de un funcional J cuadrático-elíptico: $J : \mathbb{R}^2 \longrightarrow \mathbb{R}$ de la forma

$$J(v_1, v_2) = \frac{1}{2}\left(\alpha_1 v_1^2 + \alpha_2 v_2^2\right), 0 < \alpha_1 < \alpha_2$$

para los cuales

$$J(0) = \inf_{\mathbf{v}\in\mathbb{R}^2} J(\mathbf{v})$$

y se supone que e aplica el método de gradiente de paso óptimo para resolver el problema de optimización correspondiente. Entonces, a menos que el vector inicial $\mathbf{u}_0 = (u_1^0, u_2^0)$ tiene alguno de sus componentes nulos (en cuyo caso el método converge en una iteración), el método nunca converge en un número finito de iteraciones(figura 8.5-1). Se nota que de hecho, si $\nabla J(\mathbf{u}_k) \neq 0$, es decir, si $\mathbf{u}_k = \left(u_1^k, u_2^k\right) \neq 0$, una condición necesaria y suficiente para que el punto $\mathbf{u}_{k+1}$ sea la solución del problema es que la linea $\{\mathbf{u}_k - \varrho\nabla J(\mathbf{u}_k) ; \varrho \in \mathbb{R}\}$ pase por el origen, es decir, existe un número ϱ tal que

$$u_1^k = \varrho\alpha_1 u_1^k \text{ y } u_2^k = \varrho\alpha_2 u_2^k$$

lo cual solo es posible si uno de los dos componentes u_i^k es cero (se asume $\alpha_1 \neq \alpha_2$). Ahora un cálculo simple, usando en particular la expresión del número $\varrho(\mathbf{u}_k)$ dada en la sección anterior para cualquier función cuadrática, se muestra que

$$u_1^{k+1} = \frac{\alpha_2^2(\alpha_2 - \alpha_1)u_1^k\left(u_2^k\right)^2}{\alpha_2^3\left(u_1^k\right)^2 + \alpha_2^3\left(u_2^k\right)^2}, \; u_2^{k+1} = \frac{\alpha_1^2(\alpha_1 - \alpha_2)u_2^k\left(u_1^k\right)^2}{\alpha_1^3\left(u_1^k\right)^2 + \alpha_2^3\left(u_2^k\right)^2}$$

de manera que

$$u_1^0 \neq 0 \text{ y } u_2^0 \neq 0 \Rightarrow u_1^k \neq 0 \text{ y } u_2^k \neq 0 \text{ para todo entero } k$$

¿Cómo elegir la dirección de descenso?. Asumir que el punto $\mathbf{u}_0$ no pertenece a uno de los ejes de coordenadas, y asumir el punto $\mathbf{u}_1$, construido por el método del gradiente de paso óptimo, es decir:

$$\mathbf{u}_1 = \mathbf{u}_0 - \frac{\| \mathbf{d}_0 \|^2}{\langle A\mathbf{d}_0, \mathbf{d}_0\rangle}\mathbf{d}_0 \text{ con } \mathbf{d}_0 = \nabla J(\mathbf{u}_0) \text{ y } A = \text{diag}(\alpha_i)$$

Se observa que la dirección "óptima" de descenso $\mathbf{d}_1$, en el punto $\mathbf{u}_1$ (que no es otro que el del vector $\mathbf{u}_1$; véase la *figura 8*). verifica

$$\mathbf{d}_1 \neq 0 \text{ y } \langle A\mathbf{d}_1, \mathbf{d}_0\rangle = 0$$

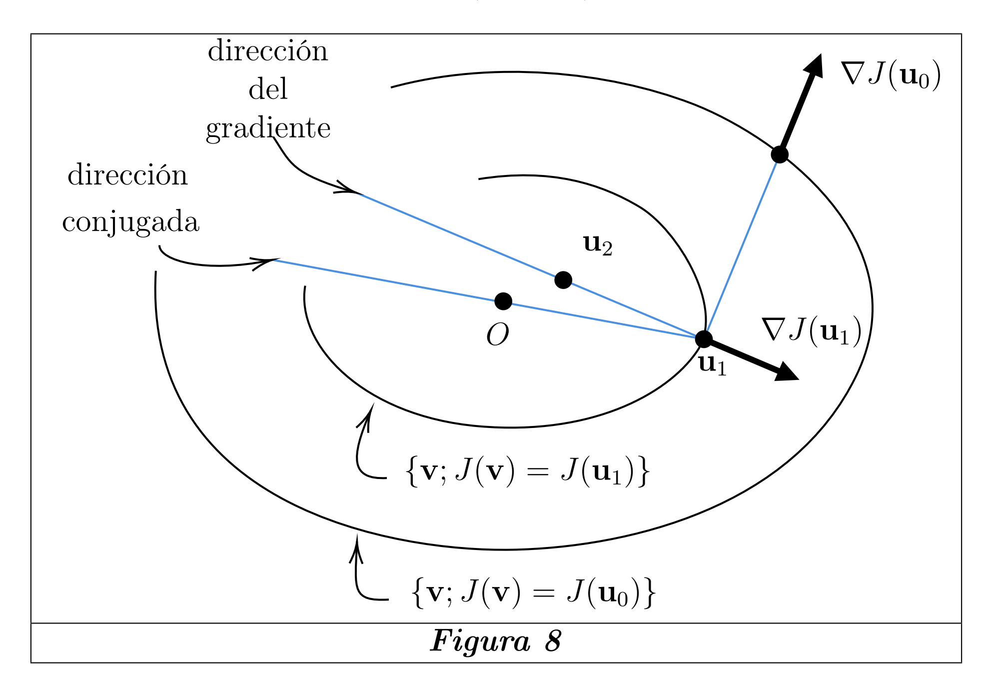


Figura 8

; estas relaciones definen de forma única la dirección del vector $\mathbf{d}_1$ (esta es la dirección "conjugada" desde la dirección de descenso anterior $\mathbf{d}_0 = \nabla J\left(\mathbf{u}_0\right)$, dependiendo del término que se especificará a continuación).
Los vectores $\nabla J\left(\mathbf{u}_0\right)$ y $\nabla J\left(\mathbf{u}_1\right)$ son linealmente independiente porque son ortogonales (véase la parte (i) del teorema 8.4-3), el punto O solución del problema también se puede considerar el mínimo del funcional en el plano que pasa por el punto $\mathbf{u}_0$, y generado por los vectores $\nabla J\left(\mathbf{u}_0\right)$ y $\nabla J\left(\mathbf{u}_1\right)$.
Es esta última idea la que se generalizará al caso de una función *cuadrática-elíptica*

$$J : \mathbf{v} \in \mathbb{R}^n \longrightarrow J(\mathbf{v}) = \frac{1}{2}\langle A\mathbf{v}, \mathbf{v}\rangle - \langle \mathbf{b}, \mathbf{v}\rangle$$

Dado un vector inicial arbitrario $\mathbf{u}_0$, suponer que los vectores $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$ ya han sido calculados. Naturalmente, se supone que

$$\nabla J(\mathbf{u}_l) \neq 0 \ , \ 0 \leq l \leq k$$

de lo contrario, el algoritmo ya está terminado. Para $l = 0, 1, \ldots, k$, se denota como G_l, el subespacio de $\mathbb{R}^n$, de dimensión $\leq l+1$, generada por los gradientes $\nabla J\left(\mathbf{u}_i\right), 0 \leq i \leq l$ (no se sabe a priori si son linealmente independientes). La idea principal del método es definir el vector "siguiente" $\mathbf{u}_{k+1}$ como el mínimo de la restricción del funcional J al conjunto

$$\begin{aligned} \mathbf{u}_{k+1} + G_k &= \{\mathbf{u}_k + \mathbf{v}_k; \mathbf{v}_k \in G_k\} \\ &= \left\{\mathbf{u}_k + \textstyle\sum_{i=0}^{k} \alpha_i \nabla J(\mathbf{u}_i); \alpha_i \in \mathbb{R}, 0 \leq i \leq k\right\} \end{aligned}$$

es decir, el punto $\mathbf{u}_{k+1}$ comprueba

$$\mathbf{u}_{k+1} \in (\mathbf{u}_k + G_k) \text{ y } J(\mathbf{u}_{k+1}) = \inf_{\mathbf{v} \in (\mathbf{u}_k + G_k)} J(\mathbf{v})$$

El conjunto $\mathbf{u}_k + G_k$ está cerrado y convexo (este es el "hiperplano" paralelo al subespacio G_k que pasa por el punto $\mathbf{u}_k$), y es funcional coercitivo y estrictamente convexo, el problema de minimización anterior admite una solución y solo una.
Por lo tanto, ya se puede prever la superioridad de este método sobre el gradiente para los que se solicita el mínimo en la única línea $\{\mathbf{u}_k - \varrho\nabla J(\mathbf{u}_k); \varrho \in \mathbb{R}\}$. Pero aún es necesario demostrar que cada uno de estos problemas de minimización en las variables k se pueden resolver de forma sencilla, lo que no es obvio a priori. Esto es sin embargo, el caso, gracias en particular a la intervención de la noción de direcciones "conjugadas" con respecto a la matriz simétrica A, como se mostrará
Las soluciones de problemas de minimización sucesivos

$$\begin{aligned}\mathbf{u}_{l+1} \in (\mathbf{u}_l + G_l) \quad \text{y} \quad J(\mathbf{u}_{l+1}) &= \text{ínf}_{\mathbf{v}\in(\mathbf{u}_l+G_l)} J(\mathbf{v}) \\ &= \text{ínf}_{\mathbf{v}\in G_l} J(\mathbf{u}_l + \mathbf{v}), 0 \le l \le k\end{aligned}$$

verificando

$$\langle \nabla J(\mathbf{u}_{l+1}), \mathbf{w}\rangle = 0 \text{ para todo } \mathbf{w} \in G_l$$

ya que los conjuntos G_l son subespacios vectoriales; en particular

$$\langle \nabla J(\mathbf{u}_{l+1}), \nabla J(\mathbf{u}_i)\rangle = 0, 0 \le i \le l \le k$$

lo que muestra que los gradientes $\nabla J(\mathbf{u}_l), 0 \le l \le k+1$ son ortogonales por pares.

∘ *Observación*

Esta propiedad es más "fuerte" que la establecida para el método de gradiente con paso óptimo, o solo los gradientes consecutivos son ortogonales.

□

Esta ortogonalidad muestra dos cosas: primero, los gradientes $\nabla J(\mathbf{u}_l), 0 \le l \le k+1$, son linealmente independientes (se suponía que eran distintos de cero); segundo, el algoritmo termina necesariamente en como máximo n iteraciones, ya que si los vectores $\nabla J(\mathbf{u}_l), 0 \le l \le n-1$, son diferentes de cero, el siguiente gradiente $\nabla J(\mathbf{u}_n)$ es forzosamente nulo (de lo contrario, se habría construido un conjunto de vectores $(n+1)$ linealmente independientes).
Se definen los $(k+1)$ vectores

$$\mathbf{u}_{l+1} - \mathbf{u}_l \overset{\text{def}}{=} \Delta_l = \sum_{i=0}^{l} \delta_i^l \nabla J(\mathbf{u}_l), 0 \le l \le k$$

y se demostrará que tienen una propiedad absolutamente notable, crucialmente relacionada al carácter cuadrático del funcional; esto de hecho hace posible escribir

$$\nabla J(\mathbf{v} - \mathbf{w}) = A(\mathbf{v} - \mathbf{w}) - \mathbf{b} = \nabla J(\mathbf{v}) + A\mathbf{w} \text{ ,para todo } \mathbf{v}, \mathbf{w} \in \mathbb{R}^n$$

y en particular

$$\nabla J(\mathbf{u}_{l+1}) = \nabla J(\mathbf{u}_l + \Delta_l) = \nabla J(\mathbf{u}_l) + A\Delta_l, 0 \le l \le k$$

De la ortogonalidad de los gradientes $\nabla J(\mathbf{u}_l), 0 \le l \le k+1$, se deduce por un lado

$$0 = \langle \nabla J(\mathbf{u}_{l+1}), \nabla J(\mathbf{u}_l)\rangle = \| \nabla J(\mathbf{u}_l) \|^2 + \langle A\Delta_l, \nabla J(\mathbf{u}_l)\rangle, 0 \le l \le k$$

y así (asumiendo $\nabla J(\mathbf{u}_l) \neq 0, 0 \le l \le k$):

$$\Delta_l \neq 0, 0 \le l \le k$$

y se deduce por otro lado para $k \geq 1$:

$$\begin{aligned} 0 = \langle \nabla J(\mathbf{u}_{l+1}), \nabla J(\mathbf{u}_l) \rangle &= \langle \nabla J(\mathbf{u}_l), \nabla J(\mathbf{u}_i) \rangle + \langle A\Delta_l, \nabla J(\mathbf{u}_i) \rangle \\ &= \langle A\Delta_l, \nabla J(\mathbf{u}_i) \rangle, 0 \leq i < l \leq k \end{aligned}$$

Dado que cada vector $\Delta_m, 0 \leq m \leq k-1$, es una combinación lineal de los vectores $\nabla J(\mathbf{u}_i), 0 \leq i \leq k-1$, se establecen las relaciones:

$$\langle A\Delta_l, \Delta_m \rangle = 0, 0 \leq m < l \leq k$$

Esto lleva a la siguiente definición: dada una matriz simétrica A, se dice que los vectores $\mathbf{w}_l, 0 \leq l \leq k$, con $k \geq 1$, son "conjugados" con respecto a la matriz A si

$$\mathbf{w}_l \neq 0, 0 \leq l \leq k \text{ , y } \langle A\mathbf{w}_l, \mathbf{w}_m \rangle = \langle A\mathbf{w}_m, \mathbf{w}_l \rangle = 0, 0 \leq m < l \leq k$$

Naturalmente, esta es una noción que involucra solo las direcciones de los vectores $\mathbf{w}_l$, que también se dice que son conjugados con respecto a la matriz A. Obsérvese también que, si la matriz A es definida positiva (como es el caso aquí), los vectores conjugados son necesariamente linealmente independientes. Efectivamente,

$$0 = \sum_{l=0}^{k} \lambda_l \mathbf{w}_l \Rightarrow 0 = \langle A \left(\sum_{l}^{k} \lambda_l \mathbf{w}_l \right), \mathbf{w}_m \rangle = \lambda_m \langle A\mathbf{w}_m, \mathbf{w}_m \rangle \Rightarrow \lambda_m = 0, 0 \leq m \leq k$$

ya que $\langle A\mathbf{w}_m, \mathbf{w}_m \rangle > 0$, de acuerdo con el carácter de la matriz A definida positiva.

∘ *Observación*

La aplicación $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^n \longrightarrow \langle A\mathbf{u}, \mathbf{v} \rangle$ es un producto escalar cuando la matriz A es simétrica definida positiva, otra forma de expresar que dos los vectores son conjugados con respecto a la matriz A, es decir que son ortogonales con respecto a este producto escalar, la ortogonalidad "habitual" correspondiente al caso parcial de la matriz de unidades.

□

Los vectores $\nabla J(\mathbf{u}_l), 0 \leq l \leq k$, y los vectores: $\Delta_l = \sum_{i=0}^{l} \delta_i^l \nabla J(\mathbf{u}_l), 0 \leq l \leq k$, siendo linealmente independiente, luego la igualdad entre matrices de orden $(k+1)$:

$$\begin{pmatrix} \Delta_0 & \Delta_1 & \cdots & \Delta_k \end{pmatrix} = \begin{pmatrix} \nabla J(\mathbf{u}_0) & \nabla J(\mathbf{u}_1) & \cdots & \nabla J(\mathbf{u}_k) \end{pmatrix} \begin{pmatrix} \delta_0^0 & \delta_0^1 & \cdots & \delta_0^k \\ & \delta_1^1 & \cdots & \delta_1^k \\ & & \ddots & \vdots \\ & & & \delta_k^k \end{pmatrix}$$

muestra que

$$\delta_l^l \neq 0, 0 \leq l \leq k$$

por lo tanto se puede escribir a priori la dirección de descenso en cada punto $\mathbf{u}_l, 0 \leq l \leq k$, en la forma

$$\mathbf{d}_l = \sum_{i=0}^{l-1} \lambda_i^l \nabla J(\mathbf{u}_i) + \nabla J(\mathbf{u}_l), 0 \leq l \leq k$$

∘ *Observación*

El descenso efectivo está en la dirección del vector $-\mathbf{d}_l$, pero por motivos de presentación, se prefiera que apareciera el signo "menos" delante del número $\varrho(\mathbf{u}_k, \mathbf{d}_k)$ introducido a continuación.

$\square$

Volviendo al cálculo del vector $\mathbf{u}_{k+1}$, se suponen que las componentes λ_l^k son conocidas; entonces se reduce a un *problema de optimización de una sola variable*: buscar $\varrho(\mathbf{u}_k, \mathbf{d}_k)$, de forma que

$$J(\mathbf{u}_k - \varrho(\mathbf{u}_k, \mathbf{d}_k)\mathbf{d}_k) = \inf_{\varrho \in \mathbb{R}} J(\mathbf{u}_k - \varrho \mathbf{d}_k)$$

y está claro que el punto $\mathbf{u}_{k+1}$ coincide entonces con el punto $\mathbf{u}_k - \varrho(\mathbf{u}_k, \mathbf{d}_k)$. De hecho, desde

$$\Delta_k = \sum_{i=0}^{k} \delta_i^k \nabla J(\mathbf{u}_i) = \delta_k^k \left\{ \sum_{i=0}^{k-1} \frac{\delta_i^k}{\delta_k^k} \nabla J(\mathbf{u}_i) + \nabla J(\mathbf{u}_k) \right\}$$

necesariamente se tiene

$$\Delta_k = \delta_k^k \mathbf{d}_k \text{ ,y } \varrho(\mathbf{u}_k, \mathbf{d}_k) = -\delta_k^k$$

Se mostrará que el cálculo efectivo de los componentes λ_i^k se realiza de una manera *notable simple*: para encontrar k ecuaciones en las k incógnitas $\lambda_k^i, 0 \leq i \leq k-1$, se escribe

$$0 = \langle A\mathbf{d}_k, \Delta_l \rangle = \langle \mathbf{d}_k, A\Delta_l \rangle = \langle \mathbf{d}_k, \nabla J(\mathbf{u}_{l+1}) + \nabla J(\mathbf{u}_l) \rangle, 0 \leq l \leq k-1$$

y de nuevo

$$\langle \sum_{i=0}^{k-1} \lambda_i^k \nabla J(\mathbf{u}_i) + \nabla J(\mathbf{u}_k), \nabla J(\mathbf{u}_{l+1}) + \nabla J(\mathbf{u}_l) \rangle = 0, 0 \leq l \leq k-1$$

Los gradientes $\nabla J(\mathbf{u}_l), 0 \leq l+1$, siendo ortogonales dos a dos, las relaciones procedentes se reducen a las ecuaciones

$$\begin{array}{ll} -\lambda_{k-1}^k \parallel \nabla J(\mathbf{u}_{k-1}) \parallel^2 + \parallel \nabla J(\mathbf{u}_k) \parallel^2 = 0 & \text{para} \quad l = k-1 \\ -\lambda_l^k \parallel \nabla J(\mathbf{u}_l) \parallel^2 + \lambda_{l+1}^k \parallel \nabla J(\mathbf{u}_{l+1}) \parallel^2 = 0 & \text{para} \quad 0 \leq l \leq k-2 \text{ si } k \geq 2 \end{array}$$

cuya solución es

$$\lambda_i^k = \frac{\parallel \nabla J(\mathbf{u}_k) \parallel^2}{\parallel \nabla J(\mathbf{u}_i) \parallel^2}, 0 \leq i \leq k-1$$

Como resultado

$$\begin{array}{rcl} \mathbf{d}_k & = & \sum_{i=0}^{k-1} \frac{\|\nabla J(\mathbf{u}_k)\|^2}{\|\nabla J(\mathbf{u}_i)\|^2} \nabla J(\mathbf{u}_i) + \nabla J(\mathbf{u}_k) \\ & = & \nabla J(\mathbf{u}_k) + \frac{\|\nabla J(\mathbf{u}_k)\|^2}{\|\nabla J(\mathbf{u}_{k-1})\|^2} \left\{ \sum_{i=0}^{k-2} \frac{\|\nabla J(\mathbf{u}_k)\|^2}{\|\nabla J(\mathbf{u}_i)\|^2} \nabla J(\mathbf{u}_i) + \nabla J(\mathbf{u}_{k-1}) \right\} \\ & = & \nabla J(\mathbf{u}_k) + \frac{\|\nabla J(\mathbf{u}_k)\|^2}{\|\nabla J(\mathbf{u}_i)\|^2} \mathbf{d}_{k-1} \end{array}$$

que proporciona un método muy simple de calcular las direcciones sucesivas de descenso, concretamente

$$\begin{cases} \mathbf{d}_0 = \nabla J(\mathbf{u}_0) \\ \mathbf{d}_l = \nabla J(\mathbf{u}_l) + \frac{\|\nabla J(\mathbf{u}_l)\|^2}{\|\nabla J(\mathbf{u}_{l-1})\|^2} \mathbf{d}_{l-1} \quad , 0 \leq l \leq k \end{cases}$$

Queda por determinar el número $\varrho(\mathbf{u}_k, \mathbf{d}_k)$ que, recordar, está definido por la relación

$$J(\mathbf{u}_k - \varrho(\mathbf{u}_k, \mathbf{d}_k)\mathbf{d}_k) = \inf_{\varrho \in \mathbb{R}} J(\mathbf{u}_k - \varrho \mathbf{d}_k)$$

Dado que la función es cuadrática, la función a minimizar es un trinomio de segundo grado:

$$\varrho \in \mathbb{R} \longrightarrow \frac{\varrho^2}{2}\langle A\mathbf{d}_k, \mathbf{d}_k\rangle - \varrho(\nabla J(\mathbf{u}_k), \mathbf{d}_k) + J(\mathbf{u}_k)$$

Por lo tanto, basta con cancelar la derivada de este trinomio, que da:

$$\varrho(\mathbf{u}_k, \mathbf{d}_k) = \frac{\langle \nabla J(\mathbf{u}_k), \mathbf{d}_k\rangle}{\langle A\mathbf{d}_k, \mathbf{d}_k\rangle}$$

Ahora se tienen todos los elementos necesarios para definir un algoritmo minimización de una función *cuadrática-elíptica*:

$$J : \mathbf{v} \in \mathbb{R}^n \longrightarrow J(\mathbf{v}) = \frac{1}{2}\langle A\mathbf{v}, \mathbf{v}\rangle - \langle \mathbf{b}, \mathbf{v}\rangle$$

llamado *método de gradiente conjugado*: a partir de un vector inicial arbitrario $\mathbf{u}_0$, se establece

$$\mathbf{d}_0 = \nabla J(\mathbf{u}_0)$$

Si $\mathbf{d}_0 = \nabla J(\mathbf{u}_0) = 0$, el algoritmo está terminado. De lo contrario, se define el número

$$r_0 = \frac{\langle \nabla J(\mathbf{u}_0), \mathbf{d}_0\rangle}{\langle A\mathbf{d}_0, \mathbf{d}_0\rangle}$$

(¡la distinción entre las dos notaciones $\mathbf{d}_0$ y $\nabla J(\mathbf{u}_0)$ es obviamente artificial en este punto!), luego el vector

$$\mathbf{u}_1 = \mathbf{u}_0 - r_0\mathbf{d}_0$$

Suponiendo que los vectores $\mathbf{u}_1, \mathbf{d}_1, \ldots, \mathbf{u}_{k-1}, \mathbf{d}_{k-1}, \mathbf{u}_k$ se construyen paso a paso, lo que implica que los gradientes $\nabla J(\mathbf{u}_l), 0 \leq l \leq k-1$, son todos diferentes de cero, pueden darse dos casos: o $\nabla J(\mathbf{u}_k) = 0$ y el algoritmo está terminado; o bien $\nabla J(\mathbf{u}_k) \neq 0$, en cuyo caso se define el vector

$$\mathbf{d}_k = \nabla J(\mathbf{u}_k) + \frac{\| \nabla J(\mathbf{u}_k) \|^2}{\| \nabla J(\mathbf{u}_{k-1}) \|^2}\mathbf{d}_{k-1}$$

entonces el número

$$r_k = \frac{\langle \nabla J(\mathbf{u}_k), \mathbf{d}_k\rangle}{\langle A\mathbf{d}_k, \mathbf{d}_k\rangle}$$

entonces el vector

$$\mathbf{u}_{k+1} = \mathbf{u}_k - r_k\mathbf{d}_k$$

y así sucesivamente.

• ***Teorema 11 :*** El método de gradiente conjugado aplicado a un funcional *cuadrático-elíptico* converge en n iteraciones como máximo.

Por lo tanto, se ha construido un nuevo método para resolver un sistema de matriz lineal simétrico definida positiva (fue, por cierto, concebido originalmente como un método de resolución de un sistema lineal), y este es un método directo, ya que conduce a la solución exacta después de un número finito de operaciones elementales. Hay que hacer la cuenta operaciones necesarias para una iteración:

(i) El cálculo de los productos escalares $\| \nabla J(\mathbf{u}_k) \|^2, \langle \nabla J(\mathbf{u}_k), \mathbf{d}_k\rangle, \langle A\mathbf{d}_k, \mathbf{d}_k\rangle$ requiere $3(n-1)$ adiciones y $3n$ multiplicaciones.

(ii) El cálculo del vector $A\mathbf{d}_k$ requiere $n(n-1)$ adiciones y n^2 multiplicaciones.

(iii) El cálculo de los vectores $\mathbf{d}_k, \mathbf{u}_{k+1}$, y $\nabla J(\mathbf{u}_{k+1}) = \nabla J(\mathbf{u}_k) - r_k A\mathbf{d}_k$ requiere $3n$ adiciones, $3n$ multiplicaciones, y 2 divisiones (para el cálculo de los cocientes $\frac{\|\nabla J(\mathbf{u}_k)\|^2}{\|\nabla J(\mathbf{u}_{k-1})\|^2}$ y r_k)

Al final, el método de gradiente conjugado requiere por lo tanto del orden de

$$\begin{cases} n^3 & \text{adiciones} \\ n^3 & \text{multiplicaciones} \\ 2n & \text{divisiones} \end{cases}$$

es decir, operaciones más elementales que el *método de Cholesky*; esto es tanto más cierto en cuanto que la presencia inevitable de errores el redondeo de cálculos prácticos que a veces conduce a continuar el proceso más allá de n iteraciones teóricamente predichas. El método del gradiente conjugado no parece ser el mejor para matrices completas (aunque disfruta de una “estabilidad numérica” que a veces es muy bienvenida), por otra parte, presenta ventajas evidentes cuando se aplica a *matrices huecas*, cuyo cálculo se prefiere a menudo evitar. De hecho, la revisión de las fórmulas de recurrencia muestra que la matriz A interviene solo por medio de cálculos de $A\mathbf{d}_k$. Este cálculo, que es más costoso cuando la matriz A está llena, es muy sencilla para ciertas matrices huecas, y en particular las resultantes de la discretización de problemas de frontera por métodos de diferencias finitas o elementos finitos: Se tiene por ejemplo que en dimensión uno, las componentes del vector A son de la forma

$$(A\mathbf{v})_i = a\mathbf{v}_{i-1} + 2b\mathbf{v}_i + a\mathbf{v}_{i+1} \ , \ \mathbf{v}_0 = \mathbf{v}_{n+1} = 0;$$

de la misma manera, fórmulas de recurrencia similares (pero un poco más elaboradas, lo cual es normal), no son difíciles de encontrar en la dimensión dos o tres. Finalmente, sucede con frecuencia en este tipo de aplicaciones que la convergencia del método es lo suficientemente rápida para permitir una reducción drástica en el número n de iteraciones esperadas teóricamente.

Con el fin de adaptar el método de gradiente conjugado a un funcional no necesariamente cuadrático, se observa que la ortogonalidad de los gradientes $\nabla J(\mathbf{u}_k)$ sucesivos, permite escribir

$$\begin{aligned} \mathbf{d}_k &= \nabla J(\mathbf{u}_k) + \frac{\|\nabla J(\mathbf{u}_k)\|^2}{\|\nabla J(\mathbf{u}_{k-1})\|^2}\mathbf{d}_{k-1} \\ &= \nabla J(\mathbf{u}_k) + \frac{\langle \nabla J(\mathbf{u}_k), \nabla J(\mathbf{u}_k) - \nabla J(\mathbf{u}_{k-1})\rangle}{\|\nabla J(\mathbf{u}_{k-1})\|^2}\mathbf{d}_{k-1} \end{aligned}$$

Es esta última expresión de la dirección de descenso la que se utiliza para definir el *método del gradiente conjugado de Polak-Ribière* para cualquier funcional J: a partir de un vector inicial arbitrario $\mathbf{u}_0$, se supone que los vectores $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$ están construidos, lo que implica que los gradientes $\nabla J(\mathbf{u}_l), 0 \leq l \leq k$, son todos diferentes de cero. Entonces se pueden presentar dos casos: $\nabla J(\mathbf{u}_k) = 0$ y el algoritmo está terminado, o bien $\nabla J(\mathbf{u}_k) \neq 0$, en cuyo caso el vector $\mathbf{u}_{k+1}$ está definido (si existe y si es único) a través de las relaciones

$$\mathbf{u}_{k+1} = \mathbf{u}_k - r_k\mathbf{d}_k \text{ ,y } J(\mathbf{u}_{k+1}) = \inf_{r\in\mathbb{R}} J(\mathbf{u}_k - r\mathbf{d}_k)$$

las direcciones de descenso sucesivas $\mathbf{d}_l$ se definen por la relación de recurrencia

$$\begin{aligned} \mathbf{d}_0 &= \nabla J(\mathbf{u}_0) \\ \mathbf{d}_l &= \nabla J(\mathbf{u}_l) + \frac{\langle \nabla J(\mathbf{u}_l), \nabla J(\mathbf{u}_l) - \nabla J(\mathbf{u}_{l-1})\rangle}{\|\nabla J(\mathbf{u}_{l-1})\|^2}\mathbf{d}_{l-1}, 1 \leq l \leq k \end{aligned}$$

∘ *Observaciones*

(*i*) Habría sido igualmente concebible a priori adaptarse al caso general el método de gradiente conjugado en su primera forma; esta adaptación lleva el nombre del *método del gradiente conjugado de Fletcher-Reeves*. El de método de Polak-Ribière sin embargo, resulta ser más eficaz en la práctica.

(*ii*) Cuando el funcional es cualquiera, no hay razón para que los gradientes $\nabla J(\mathbf{u}_k)$ obtenidos por el método Polak-Ribière siguan siendo ortogonales de dos a dos y así para que el algoritmo termine en un número finito de iteraciones.

(*iii*) Por construcción, el método Polak-Robière coincide con el de Fletcher-Reeves cuando se aplica a un funcional cuadrático.

□

7.3. Métodos de relación, gradiente y de penalización para problemas con restricciones

En esta sección, se estudiarán en los problemas con restricciones que se presentan de la siguiente manera: dada un conjunto U de un espacio vectorial V y un funcional $J : V \longrightarrow \mathbb{R}$, encontrar $\mathbf{u}$ tal que

$$\mathbf{u} \in U \subseteq \mathbb{R}^n, \text{ y } J(\mathbf{u}) = \inf_{\mathbf{v}\in U} J(\mathbf{v}) \tag{P}$$

Es hora de ampliar la definición del método de relajación a problemas con restricciones para las que el conjunto U es de la forma particular

$$U = \{\mathbf{v} = (v_i) \in \mathbb{R}^n; a_i \leq v_i \leq b_i, 1 \leq i \leq n\} = \prod_{i=1}^{n} [a_i, b_i]$$

sin excluir los casos $a_i = -\infty$ y/o $b_i = +\infty$. Conociendo el vector $\mathbf{u}_k = \left(u_i^k\right)_{i=1}^n$, se define el vector $\mathbf{u}_{k+1} = \left(u_i^{k+1}\right)_{i=1}^n$ resolviendo sucesivamente los n problemas de minimización de una variable:

$$\begin{array}{lcl}
J\left(\left[u_1^{k+1}\right], u_2^k, u_3^k, \quad \ldots \quad, u_{n-1}^k, u_n^k\right) & = & \inf_{a_1\leq\zeta\leq b_1} J\left(\zeta, u_2^k, u_3^k, \ldots, u_n^k\right) \\
J\left(u_1^{k+1}, \left[u_2^{k+1}\right], u_3^k, \quad \ldots \quad, u_{n-1}^k, u_n^k\right) & = & \inf_{a_2\leq\zeta\leq b_2} J\left(u_1^{k+1}, \zeta, u_3^k, \ldots, u_n^k\right) \\
 & \vdots & \\
J\left(u_1^{k+1}, u_2^{k+1}, u_3^{k+1}, \quad \ldots \quad, u_{n-1}^{k+1}, \left[u_n^{k+1}\right]\right) & = & \inf_{a_n\leq\zeta\leq b_n\mathbb{R}} J\left(u_1^{k+1}, u_2^{k+1}, u_3^{k+1},\right. \\
 & & \left.\ldots, u_{n-1}^{k+1}, \zeta\right)
\end{array}$$

• ***Teorema 12 :*** Si el funcional $J : \mathbb{R}^n \longrightarrow \mathbb{R}$ es elíptico y el conjunto U tiene la forma:

$$U = \prod_{i=1}^{n} [a_i, b_i] \text{ , sin excluir } a_i = -\infty \text{ y/o } b_i = +\infty$$

, el método de relajación converge

∘ *Demostración*

Se sigue del *teorema 8*, la única novedad es el reemplazo de las caracterizaciones $\partial_l J(\mathbf{u}_{k;l}) = 0, 1 \leq l \leq n$, y $\nabla J(\mathbf{u}) = 0$ del caso sin restricciones por los requisitos

necesarios y suficientes de minimización:

$$\begin{cases} \partial_l J(\mathbf{u}_{k;l}) \left(v_l - u_l^{k+1}\right) \geq 0 \text{ para todo } v_l \in [a_l, b_l], 1 \leq l \leq n \\ \langle \nabla J(\mathbf{v}), \mathbf{v} - \mathbf{u} \rangle \geq 0 \text{ para todo } \mathbf{v} \in U \end{cases}$$

De hecho, se verifican las desigualdades

$$J(\mathbf{u}_{k;l-1}) - J(\mathbf{u}_{k;l}) \geq \frac{\alpha}{2} \parallel \mathbf{u}_{k;l-1} - \mathbf{u}_{k;l} \parallel^2$$
$$\alpha \parallel \mathbf{u}_{k+1} - \mathbf{u} \parallel^2 \leq \langle \nabla J(\mathbf{u}_{k+1}), \mathbf{u}_{k+1} - \mathbf{u} \rangle$$

obtenidas respectivamente en los pasos (i) y (iii) de la demostración del teorema antes mencionado.

■

○ *Observación*

No es posible extender el método de relajación sin cuidado a conjuntos de U más generales; por ejemplo, si

$$J(\mathbf{v}) = \left(v_1^2, v_2^2\right) \text{ y } U = \left\{\mathbf{v} = (v_1, v_2) \in \mathbb{R}^2; v_1 + v_2 \geq 2\right\}$$

se convence fácilmente (*figura 9*) de que, a menos que uno de los componentes del vector $\mathbf{u}_0$ inicial, es 1, el algoritmo definido por

$$\begin{aligned} J\left(u_1^{k+1}, u_2^k\right) &= \text{ínf}_{\xi \geq 2 - u_1^k} J\left(\zeta, u_2^k\right) \\ J\left(u_1^{k+1}, u_2^{k+1}\right) &= \text{ínf}_{\zeta \geq 2 - u_2^k} J\left(u_1^{k+1}, \xi\right) \end{aligned}$$

se bloquea en el límite del conjunto U.

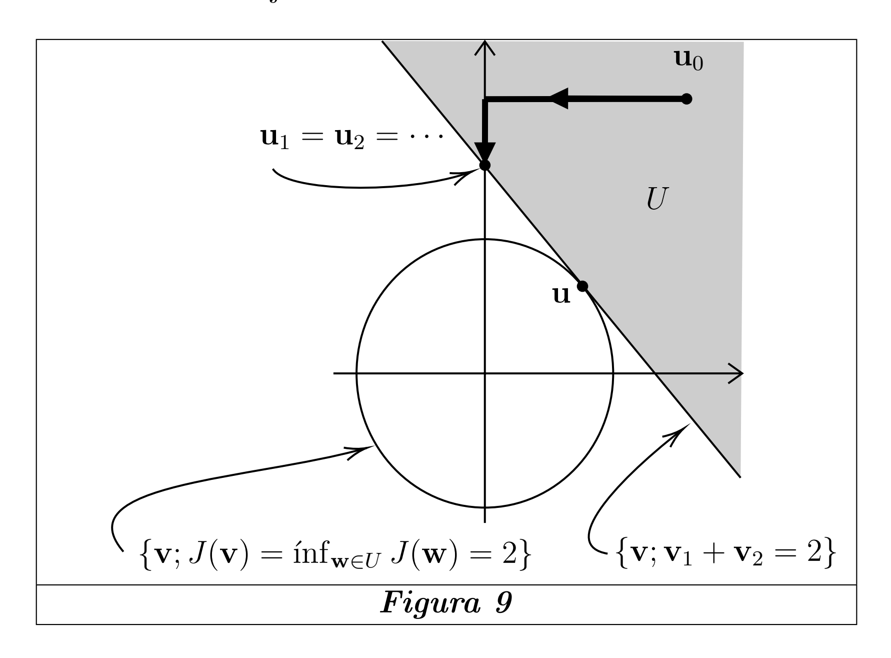


Figura 9

□

Ahora considerar el problema (P) asociado con un conjunto convexo U y un funcional convexo. Un elemento $\mathbf{u} \in U$ es entonces la solución de la problema (P) si comprueba las siguientes condiciones necesarias y suficientes:

$$\langle \nabla J(\mathbf{v}), \mathbf{v} - \mathbf{u} \rangle \geq 0 \text{ para todo } \mathbf{v} \in U$$

Uno no puede dejar de notar la analogía entre estas condiciones y la caracterización (*teorema 1*)

$$\langle \mathbf{u} - \mathbf{w}, \mathbf{v} - \mathbf{u} \rangle \geq 0 \text{ para todo } \mathbf{v} \in U$$

de la proyección $\mathbf{u}$ de un elemento $\mathbf{w}$ de un espacio de Hilbert V en un subconjunto $U \subset V$ no vacío, convexo y cerrado. Más precisamente, designando por P el operador de proyección del espacio V en el conjunto U, se tienen las siguientes equivalencias:

$$\begin{aligned} \mathbf{u} \in U \quad \text{y} \quad J(\mathbf{u}) = \text{ínf}_{\mathbf{v}\in U} J(\mathbf{v}) &\Leftrightarrow \mathbf{u} \in U \text{ y } \langle \nabla J(\mathbf{u}), \mathbf{v} - \mathbf{u} \rangle \geq 0 \text{ para todo } \mathbf{v} \in U \\ &\Leftrightarrow \mathbf{u} \in U \text{ y } \langle \mathbf{u} - \{\mathbf{u} - \varrho \nabla J(\mathbf{u})\}, \mathbf{v} - \mathbf{u} \rangle \geq 0 \text{ para todo } \mathbf{v} \in U, \varrho > 0 \\ &\Leftrightarrow \mathbf{u} = P(\mathbf{u} - \varrho \nabla J(\mathbf{u})) \text{ para todo } \varrho < 0 \end{aligned}$$

En otras palabras, la solución $\mathbf{u}$ aparece, para cualquier $\varrho > 0$, como un punto fijo de la aplicación

$$g : \mathbf{v} \in V \longrightarrow g(\mathbf{v}) = P(\mathbf{v} - \varrho \nabla J(\mathbf{v})) \in U \subset V$$

Por lo tanto, es natural definir como un método de aproximación de la solución de la problema (P) el método de aproximaciones sucesivas aplicado a g: Dado un elemento arbitrario $\mathbf{u}_0 \in V$, se define la sucesión $(\mathbf{u}_k)_{k\geq 0}$ por:

$$\mathbf{u}_{k+1} = g(\mathbf{u}_k) = P(\mathbf{u}_k - \varrho \nabla J(\mathbf{u}_k)), k \geq 0$$

En el caso de que $U = V$, el operador de proyección P es la identidad, y la relación anterior se reduce a

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \varrho \nabla J(\mathbf{u}_k), k \geq 0$$

Por lo tanto, se encuentra el método de gradiente de paso fijo para un problema sin restricciones, que se estudia en la sección 8.4. Esta es la razón por la que el método que se acaba de describir se llama el *método de gradiente con proyección de paso fijo*.
Para demostrar su convergencia, basta simplemente con verificar que, si el parámetro $\varrho > 0$ ha sido elegido adecuadamente, entonces la aplicación $g : V \longrightarrow V$ es una contracción, es decir, existe un número β tal que

$$\beta < 1 \text{ ,y } \| g(\mathbf{v}_1) - g(\mathbf{v}_2) \| \leq \beta \| \mathbf{v}_1 - \mathbf{v}_2 \| \text{ para todo } \mathbf{v}_1, \mathbf{v}_2 \in V$$

De hecho, esta hipótesis conduce a la existencia de un punto fijo y a la convergencia del método aproximaciones sucesivas tan pronto como se supone que el espacio V está completo; por eso la compacidad no interviene en la demostración. Porque no introduce ninguna dificultad adicional, incluso se considera el *método de gradiente (más general) con proyección de paso variable*, definida por

$$\mathbf{u}_{k+1} = P(\mathbf{u}_k - \varrho_k \nabla J(\mathbf{u}_k)), \varrho_k > 0, k \geq 0$$

• ***Teorema 13:*** Sean V un espacio de Hilbert, U un subconjunto no vacío, convexo y cerrado de V y $J : V \longrightarrow \mathbb{R}$ un funcional derivable sobre V, se supone que existen sos variables α y M tales que:

$$\alpha > 0 \text{ y } \langle \nabla J(\mathbf{v}) - \nabla J(\mathbf{u}), \mathbf{v}-\mathbf{u}\rangle \geq \alpha \left\|\mathbf{v}-\mathbf{u}\right\|^2 \text{ para todo } \mathbf{u}, \mathbf{v} \in V$$

$$\left\|\nabla J(\mathbf{v}) - \nabla J(\mathbf{u})\right\| \leq M \left\|\mathbf{v}-\mathbf{u}\right\|^2 \text{ para todo } \mathbf{u}, \mathbf{v} \in V$$

Si existen dos números a y b tales que:

$$0 < a \leq \varrho_k \leq b < \frac{2\alpha}{M^2} \text{ para todo entero } k \geq 0$$

, el método de gradiente con restricciones converge y la convergencia es geométrica si existe una constante $\beta = (\alpha, M, a, b)$ tal que:

$$\beta < 1 \text{ y } \left\|\mathbf{u}_k - \mathbf{u}\right\| \leq \beta^k \left\|\mathbf{u}_0 - \mathbf{u}\right\|$$

○ *Demostración*

Para cualquier entero $k \geq 0$, se define la aplicación

$$g_k : \mathbf{v} \in V \longrightarrow g_k(\mathbf{v}) = P(\mathbf{v} - \varrho_k \nabla J(\mathbf{v})) \in U \subset V$$

Debido al hecho de que la proyección “no aumenta las distancias” (*teorema 1*), y con las suposiciones hechas sobre el funcional, se deducen las desigualdades

$$\begin{aligned} \| g_k(\mathbf{v}_1) - g_k(\mathbf{v}_2) \|^2 &= \| P(\mathbf{v}_1 - \varrho_k \nabla J(\mathbf{v}_1)) - P(\mathbf{v}_2 - \varrho_k \nabla J(\mathbf{v}_2)) \|^2 \\ &\leq \| (\mathbf{v}_1 - \mathbf{v}_2) - \varrho_k(\nabla J(\mathbf{v}_1) - \nabla J(\mathbf{v}_2)) \|^2 \\ &= \| \mathbf{v}_1 - \mathbf{v}_2 \|^2 - 2\varrho_k \langle \nabla J(\mathbf{v}_1) - \nabla J(\mathbf{v}_2), \mathbf{v}_1 - \mathbf{v}_2 \rangle \\ &\quad + \varrho_k^2 \| \nabla J(\mathbf{v}_1) - \nabla J(\mathbf{v}_2) \|^2 \\ &\leq (1 - 2\alpha\varrho_k + M^2\varrho_k^2) \| \mathbf{v}_1 - \mathbf{v}_2 \|^2 \end{aligned}$$

asumiendo $\varrho_k > 0$. Por cierto, ya se estableció (en la demostración del *teorema 10*) la existencia de una constante $\beta = \beta(\alpha, M, a, b)$ tal que

$$\left(1 - 2\alpha\varrho_k + M^2\varrho_k^2\right)^{\frac{1}{2}} \leq \beta < 1 \text{ para todo } k \geq 0$$

cuando los números a y b verifican las suposiciones de la declaración. Dado que la solución $\mathbf{u}$ del problema (P) es un punto fijo de cada aplicación ϱ_k, se puede escribir

$$\| \mathbf{u}_{k+1} - \mathbf{u} \| = \| g_k(\mathbf{u}_k) - g_k(\mathbf{u}) \| \leq \beta \| \mathbf{u}_{k+1} - \mathbf{u} \|$$

y se demuestra la convergencia geométrica.

■

○ *Observaciones*

(*i*) La existencia del punto fijo de la aplicación $g(\mathbf{v}) = P(\mathbf{V} - \nabla J(\mathbf{v}))$ asociado al método de gradiente con proyección de paso fijo, y por lo tanto la existencia de una solución $\mathbf{u}$ de las desigualdades $\langle \nabla J(\mathbf{u}, \mathbf{v}-\mathbf{u})\rangle$ para cualquier $\mathbf{v} \in U$ proporciona una prueba de la existencia de una solución del problema (P) asociado con un conjunto U y un funcional verificando las hipótesis del teorema actual, que aparece como un caso especial del resultado de la *teorema 5*.

(*ii*) Si $U = V$, encontramos la convergencia del método de gradiente de paso variable, ya establecido en el *teorema 10.*

(*iii*) En el caso de una función cuadrática elíptica

$$J : \mathbf{v} \in \mathbb{R}^n \longrightarrow J(\mathbf{v}) = \frac{1}{2}\langle A\mathbf{v}, \mathbf{v}\rangle - \langle \mathbf{b}, \mathbf{v}\rangle, A = A^t$$

se puede mostrar, exactamente como en el caso sin restricciones (*teorema 10*), que la convergencia geométrica tiene lugar para $g_k \in \left[a, \tilde{b}\right] \subset \left]0, \frac{2}{\lambda_n}\right[$, mientras que, en este caso particular, el teorema anterior proporciona solo convergencia para $g_k \in [a, b] \subset \left]0, \frac{2\lambda_1}{\lambda_n^2}\right[$ (se recuerda que λ_1 y λ_n los valores propios extremos de la matriz A).

□

Por lo tanto, en principio, los métodos de gradiente con proyección proporcionan métodos de aproximaciones aplicables a una amplia clase de problemas de programación convexos, pero esto es un señuelo desde el punto de vista "numérico", por la sencilla razón de que el operador proyección sobre cualquier subconjunto convexo y cerrado no se conoce explícitamente en general.

Una excepción notable son los subconjuntos U de $V = \mathbb{R}^n$ de la forma $\prod_{i=1}^{n}[a_i, b_i] \subset V = \mathbb{R}^n$, para lo cual ya se ha incorporado en la sección 8.1 el operador de proyección asociado. Por ejemplo, si

$$U = \mathbb{R}^n_+ = \{\mathbf{v} \ \in \mathbb{R}^n ; \mathbf{v} \geq 0\}$$

y si este conjunto U está asociado con una función cuadrática elíptica

$$J : \mathbf{v} \in \mathbb{R}^n \longrightarrow J(\mathbf{v}) = \frac{1}{2}\langle A\mathbf{v}, \mathbf{v}\rangle - \langle \mathbf{b}, \mathbf{v}\rangle$$

el vector $\mathbf{u}_{k+1} = \left(u_i^{k+1}\right)_{i=1}^n$ está calculado a partir del vector $\mathbf{u}_k = \left(u_i^k\right)_{i=1}^n$ por las relaciones

$$u_i^{k+1} = \text{máx}\left\{u_i^k - \varrho_k(A\mathbf{u}_k - \mathbf{b})_i, 0\right\}, 1 \leq i \leq n$$

Con la excepción de estos casos especiales, los problemas con restricciones deben ser procesados por otros métodos. Este es el caso, en particular, de los métodos de penalización, el principio de que se basa en el siguiente resultado:

• ***Teorema 14:*** Sean $J : \mathbb{R}^n \longrightarrow \mathbb{R}$ una función continua, coercitiva y estrictamente convexa, U un subconjunto no vacío, convexo y cerrado de $\mathbb{R}^n$, y $\psi : \mathbb{R}^n \longrightarrow \mathbb{R}$ una función continua y convexa que verifica

$$\psi(\mathbf{v}) \geq 0 \text{ para todo } \mathbf{v} \in \mathbb{R}^n \text{ y } \psi(\mathbf{v}) = 0 \Leftrightarrow \mathbf{v} \in U$$

Entonces, para cada $\varepsilon > 0$, existe uno y solo un elemento $\mathbf{u}_\varepsilon$ que satisface

$$\mathbf{u}_\varepsilon \in \mathbb{R}^n \text{ y } J_\varepsilon\left(\mathbf{u}_\varepsilon\right) = \inf_{\mathbf{v} \in \mathbb{R}^n} J_\varepsilon\left(\mathbf{v}\right) \text{ donde } J_\varepsilon\left(\mathbf{v}\right) \stackrel{\text{def}}{=} J\left(\mathbf{v}\right) + \frac{1}{\varepsilon}\psi(\mathbf{v})$$

y $\text{lím}_{\varepsilon \to 0} \mathbf{u}_\varepsilon = \mathbf{u}$, donde $\mathbf{u}$ es la solución del problema (P)

∘ *Demostración*

Está claro que en el problema (P) y cada problema (P_ε) hay una solución y solo una. Los funcionales J, de hecho, siguen siendo coercitivos (ya que $J_\varepsilon(\mathbf{u}) \geq J(\mathbf{v})$) y estrictamente convexos (ya que la suma de una función estrictamente convexa y una función convexa es estrictamente convexa). Como

$$J(\mathbf{u}_\varepsilon) \leq J(\mathbf{u}_\varepsilon) + \frac{1}{\varepsilon}\psi(\mathbf{u}_\varepsilon) = J_\varepsilon(\mathbf{u}_\varepsilon) \leq J_\varepsilon(\mathbf{u}) = J(\mathbf{u})$$

se deduce de la coercitividad del funcional J que la familia $(\mathbf{u}_\varepsilon)_{\varepsilon>0}$ está acotada.
Por compacidad, existe una sucesión extraída $(\mathbf{u}_{\varepsilon'})_{\varepsilon'>0}$ y un elemento $\mathbf{u}' \in \mathbb{R}^n$ tal que

$$\lim_{\varepsilon' \to 0} \mathbf{u}_{\varepsilon'} = \mathbf{u}'$$

De las desigualdades $J(\mathbf{u}_{\varepsilon'}) \leq J(\mathbf{u})$ y de la continuidad de la función J, se deduce

$$J(\mathbf{u}') = \lim_{\varepsilon' \to 0} J(\mathbf{u}_{\varepsilon'}) \leq J(\mathbf{u})$$

Ya que

$$0 \leq \psi(\mathbf{u}_{\varepsilon'}) \leq \varepsilon'(J(\mathbf{u}) - J(\mathbf{u}_{\varepsilon'}))$$

y como la sucesión $(\mathbf{u}_{\varepsilon'})_{\varepsilon'>0}$ converge, los números $\{J(\mathbf{u}) - J(\mathbf{u}_{\varepsilon'})\}$ se incrementan de forma independiente de $\varepsilon\prime$; por lo tanto,

$$0 = \lim_{\varepsilon' \to 0} \psi(\mathbf{u}_{\varepsilon'}) = \psi(\mathbf{u}')$$

puesto que la función ψ es continua, lo que muestra que $\mathbf{u}' \in U$ y, por lo tanto, que $\mathbf{u} = \mathbf{u}'$ ya que $J(\mathbf{u}_{\varepsilon'}) \leq J(\mathbf{u})$ y $\mathbf{u}$ es la única solución del problema (P). La singularidad de esta solución muestra también que toda la familia $(\mathbf{u}_\varepsilon)_{\varepsilon>0}$ converge al elemento $\mathbf{u}$ (de hecho, se puede reproducir el razonamiento anterior para *todas* las sucesiones extraídas).

■

◦ *Observación*

Se muestra que cualquier función convexa $\varphi : \mathbb{R}^n \longrightarrow \mathbb{R}$ es necesariamente continua; por lo tanto, esta "hipótesis" es superflua.

□

Como aplicación, se considera el problema de programación convexa: Dado un funcional $J : \mathbb{R}^n \longrightarrow \mathbb{R}$ estrictamente convexo y funciones $\varphi_i : \mathbb{R}^n \longrightarrow \mathbb{R}, 1 \leq i \leq m$, convexas, encontrar $\mathbf{u}$ tal que

$$\mathbf{u} \in U = \{\mathbf{v} \in \mathbb{R}^n; \varphi_i(\mathbf{v}) \leq 0, 1 \leq i \leq m\}, J(\mathbf{u}) = \inf_{\mathbf{v} \in U} J(\mathbf{v})$$

Como la función ψ satisface las hipótesis del teorema 8.3-3, se toma por ejemplo

$$\psi : \mathbf{v} \in \mathbb{R}^n \longrightarrow \psi(\mathbf{v}) = \sum_{i=1}^{m} \max\{\varphi_i(\mathbf{v}), 0\}$$

Por lo tanto, el propósito esencial de un método de penalización es reemplazar un problema optimización con restricciones por una serie de problemas sin restricciones (que en principio se sabe resolver), asociado con el funcional penalizado $J_\varepsilon, \varepsilon > 0$.

◦ *Observación*

El alcance práctico de los métodos de penalización está limitado por la dificultad para construir efectivamente "buenas" funciones ψ (por ejemplo, derivables, que por cierto no es el caso para el ejemplo anterior) que cumpla las condiciones del teorema.

□

Otra forma de volver a la resolución de problemas sin restricciones está relacionada con la noción de *dualidad*. Su estudio y construcción de métodos de aproximación correspondientes serían un buen tema de estudio próximo.

7.4. Ejercicios

Ejercicio 7.1: El objeto de este problema es el estudio de un método de gradiente con paso óptimo, en ausencia de hipótesis de elipticidad del funcional $J : \mathbb{R}^n \longrightarrow \mathbb{R}$. Las únicas hipótesis son las siguientes: Se supone conocido un punto $\mathbf{u}_0 \in \mathbb{R}^n$ tal que el conjunto:

$$U = \{\mathbf{v} \in \mathbb{R}^n | J(\mathbf{v}) \leq J(\mathbf{u}_0)\}$$

sea un conjunto compacto de $\mathbb{R}^n$; se asume que el funcional J es dos veces derivable en cualquier punto de U; finalmente, además se supone que existe una constante M tal que:

$$\left|\left\langle \nabla^2 J(\mathbf{v})\,\mathbf{w}, \mathbf{w}\right\rangle\right| \leq M \|\mathbf{w}\|^2, \forall \mathbf{v} \in U, \mathbf{w} \in \mathbb{R}^n$$

(i) Sea $\mathbf{v}$ un punto del conjunto U tal que $\nabla J(\mathbf{v}) \neq 0$. Demostrar que el número:

$$\tau(\mathbf{v}) = \sup\{\rho \geq 0 | \, [\mathbf{v}, \mathbf{v} - \rho \nabla J(\mathbf{v})] \subset U\}$$

es finito y estrictamente positivo

(ii) Desde el punto $\mathbf{u}_0$, se construye una sucesión de puntos $\mathbf{u}_k \in U$ de la siguiente manera:

(a) Si $\nabla J(\mathbf{u}_k) = 0$, el algoritmo está terminado

(b) Si $\nabla J(\mathbf{u}_k) \neq 0$, se elige el punto $\mathbf{u}_{k+1}$ de tal manera que:

$$(*)\begin{cases} \mathbf{u}_{k+1} \in [\mathbf{u}_k, \mathbf{u}_k - \tau(\mathbf{u}_k)\nabla J(\mathbf{u}_k)] \\ J(\mathbf{u}_{k+1}) = \text{ínf}_{0\leq\rho\leq\tau(\mathbf{u}_k)} J(\mathbf{u}_k - \rho\nabla J(\mathbf{u}_k)) \end{cases}$$

; en el caso (b), demostrar que existe al menos un punto $\mathbf{u}_{k+1}$ que verifica las relaciones $(*)$; en el caso de que tal punto no se defina de manera única, se entiende que se hace una elección arbitraria entre todos los que son posibles

(iii) Demuestre que en el caso (ii),

$$J(\mathbf{u}_k) - J(\mathbf{u}_{k+1}) \geq \frac{\|\nabla J(\mathbf{u}_k)\|^2}{2M}$$

Se supone, para las cuestiones (iv) y (v), que siempre se tiene la eventualidad (ii), que define una sucesión infinita $(\mathbf{u}_k)$

(iv) Muestre que existe una subsucesión $(\mathbf{u}_{k'})$ de $(\mathbf{u}_k)$ y un punto $\mathbf{u} \in U$ tal que:

$$\lim_{k'\to\infty} \mathbf{u}_{k'} = \mathbf{u} \text{ , y } \nabla J(\mathbf{u}) = 0$$

(*v*) Si solo hay un punto $\mathbf{u} \in U$ verificando $\nabla J(\mathbf{u}) = 0$, demuestre que el punto $\mathbf{u}$ es un mínimo estricto del funcional $J : U \longrightarrow \mathbb{R}$ y que toda la sucesión $(\mathbf{u}_k)$ converge a este punto

∘ *Solución*

(*i*) Se considera la *fórmula de Taylor-Young* para funciones dos veces derivables:

• ***Teorema (Fórmulas de Taylor para funciones dos veces derivables):*** Sea $f : \Omega \subset X \longrightarrow Y$ y $[a, a+h]$ cualquier segmento cerrado contenido en Ω

Fórmula de Taylor-Young: Si f es dos veces derivable en Ω, entonces

$$f(a+h) = f(a) + f'(a)h + \frac{1}{2}f''(a)(h,h) + ||h||^2\varepsilon(h), \lim_{h\to 0}\varepsilon(h) = 0$$

Aplicando la fórmula para $J(\mathbf{v} - \rho\nabla J(\mathbf{v}))$:

$$\begin{aligned} J(\mathbf{v} - \rho\nabla J(\mathbf{v})) &= J(\mathbf{v}) + J'(\mathbf{v})\rho\nabla J(\mathbf{v}) + \frac{1}{2}J''(\mathbf{v})\left(\rho\nabla J(\mathbf{v}), \rho\nabla J(\mathbf{v})\right) \\ &\quad + \|\rho\nabla J(\mathbf{v})\|^2 \varepsilon(\rho) \\ &= J(\mathbf{v}) + \langle\nabla J(\mathbf{v}), \rho\nabla J(\mathbf{v})\rangle + \frac{1}{2}\left\langle\nabla^2 J(\mathbf{v})\rho\nabla J(\mathbf{v}), \rho\nabla J(\mathbf{v})\right\rangle \\ &\quad + \rho^2\|\nabla J(\mathbf{v})\|^2 \varepsilon(\rho) \\ &= J(\mathbf{v}) + \rho\|\nabla J(\mathbf{v})\|^2 + \frac{\rho^2}{2}\left\langle\nabla^2 J(\mathbf{v})\nabla J(\mathbf{v}), \nabla J(\mathbf{v})\right\rangle \\ &\quad + \rho^2\|\nabla J(\mathbf{v})\|^2 \varepsilon(\rho) \end{aligned}$$

con $\lim_{\rho\to 0}\varepsilon(\rho) = 0$, notar además que $|\langle\nabla^2 J(\mathbf{v})\mathbf{w}, \mathbf{w}\rangle| \leq M\|\mathbf{w}\|^2, \forall\mathbf{v} \in U, \mathbf{w} \in \mathbb{R}^n$ por lo tanto para todo $\mathbf{v} \in U$ se tiene:

$$\begin{aligned} J(\mathbf{v} - \rho\nabla J(\mathbf{v})) &= J(\mathbf{v}) + \rho\|\nabla J(\mathbf{v})\|^2 + \frac{\rho^2}{2}\left\langle\nabla^2 J(\mathbf{v})\nabla J(\mathbf{v}), \nabla J(\mathbf{v})\right\rangle \\ &\quad + \rho^2\|\nabla J(\mathbf{v})\|^2 \varepsilon(\rho) \\ &\leq J(\mathbf{v}) + \rho\|\nabla J(\mathbf{v})\|^2 + \frac{\rho^2}{2}M\|\nabla J(\mathbf{v})\|^2 + \rho^2\|\nabla J(\mathbf{v})\|^2\varepsilon(\rho) \\ &\leq J(\mathbf{u}_0) + \left(\frac{M}{2}\rho^2 - \rho + \rho^2\varepsilon(\rho)\right)\|\nabla J(\mathbf{v})\|^2 \end{aligned}$$

, donde $J(\mathbf{v} - \rho\nabla J(\mathbf{v})) \leq J(\mathbf{u}_0)$ para $\rho > 0$ bastante pequeño (notar que $U = \{\mathbf{v} \in \mathbb{R}^n | J(\mathbf{v}) \leq J(\mathbf{u}_0)\}$). Por otro lado, la pertenencia de $(\mathbf{v} - \rho\nabla J(\mathbf{v}))$ en el compacto U implica usar la desigualdad triangular $\|\rho\nabla J(\mathbf{v})\| \leq \|\mathbf{v}\| + \|\rho\nabla J(\mathbf{v})\| \leq 2\sup_{\mathbf{w}\in\mathbf{U}}\|\mathbf{w}\| < +\infty$. Se deduce que para cualquier $\mathbf{v} \in U$ con $\nabla J(\mathbf{v}) \neq 0$ el número $\tau(\mathbf{v})$ es estrictamente positivo.

(*ii*) Considerando la función

$$\varphi_k : \rho \in \mathbb{R} \mapsto \varphi_k(\rho) = J(\mathbf{u}_k - \rho\nabla J(\mathbf{u}_k))$$

, al fijar algún k y dado que J es dos veces derivable, se tiene que φ_k es continua sobre $\mathbb{R}$; luego alcanza su mínimo en cualquier compacto que no esté vacío (teorema de Weierstrass), en particular en $[0, \tau(\mathbf{u}_k)]$

(iii) Hay que situarse en el caso (ii); se tiene $J(\mathbf{u}_{k+1}) < J(\mathbf{u}_k)$, y la definición del número $\tau(\mathbf{u}_k)$ implica que el punto $\mathbf{u}_{k+1}$ es un punto interior del intervalo $[\mathbf{u}_k, \mathbf{u}_k - \tau(\mathbf{u}_k)\nabla J(\mathbf{u}_k)]$. En consecuencia, denotando como ρ_k el número tal que $\mathbf{u}_{k+1} = \mathbf{u}_k - \rho_k \nabla J(\mathbf{u}_k)$, se tiene la condición necesaria de optimalidad $\varphi_k'(\rho_k) = 0$ de la que se deduce la relación de ortogonalidad:

$$
\begin{aligned}
\varphi_k'(\rho_k) &= (J(\mathbf{u}_{k+1}))' \\
&= (J(\mathbf{u}_k - \rho_k \nabla J(\mathbf{u}_k)))' \\
&= J'(\mathbf{u}_{k+1}) \nabla J(\mathbf{u}_k) \\
&= \langle \nabla J(\mathbf{u}_{k+1}), \nabla J(\mathbf{u}_k) \rangle = 0
\end{aligned}
$$

Ahora se considera la fórmula de *Taylor-Maclaurin*:

• ***Teorema (Fórmulas de Taylor para funciones dos veces derivables):*** Sea $f : \Omega \subset X \longrightarrow Y$ y $[a, a+h]$ cualquier segmento cerrado contenido en Ω

Fórmula de Taylor-Maclaurin: Si $f \in \mathcal{C}^1(\Omega)$, con f dos veces derivable en $[a, a+h]$, y $Y = \mathbb{R}$, entonces

$$f(a+h) = f(a) + f'(a)h + \frac{1}{2} f''(a+\theta h)(h,h), 0 < \theta < 1$$

De acuerdo con las fórmulas de Taylor-MaLaurin aplicadas en el punto $\mathbf{u}_k$ con el aumento $-\rho \nabla J(\mathbf{u}_k)$, y en el punto $\mathbf{u}_{k+1}$ con el aumento $+\rho \nabla J(\mathbf{u}_k)$, existen dos puntos intermedios $\mathbf{u}_k^-$ y $\mathbf{u}_{k+1}^+$ tales que

$$
\begin{aligned}
J(\mathbf{u}_{k+1}) - J(\mathbf{u}_k) &= J'(\mathbf{u}_k)(-\rho_k \nabla J(\mathbf{u}_k)) + \frac{1}{2} J''\left(\mathbf{u}_k^-\right)(-\rho_k \nabla J(\mathbf{u}_k), -\rho_k \nabla J(\mathbf{u}_k)) \\
&= J'(\mathbf{u}_k)(-\rho_k \nabla J(\mathbf{u}_k)) + \frac{1}{2} J''\left(\mathbf{u}_k^-\right)(-\rho_k \nabla J(\mathbf{u}_k), -\rho_k \nabla J(\mathbf{u}_k)) \\
&= \langle \nabla J(\mathbf{u}_k), -\rho_k \nabla J(\mathbf{u}_k) \rangle + \frac{\rho_k^2}{2} \left\langle \nabla^2 J\left(\mathbf{u}_k^-\right) \nabla J(\mathbf{u}_k), \nabla J(\mathbf{u}_k) \right\rangle \\
&= -\rho_k \|\nabla J(\mathbf{u}_k)\|^2 + \frac{\rho_k^2}{2} \left\langle \nabla^2 J\left(\mathbf{u}_k^-\right) \nabla J(\mathbf{u}_k), \nabla J(\mathbf{u}_k) \right\rangle
\end{aligned}
$$

y

$$
\begin{aligned}
J(\mathbf{u}_k) - J(\mathbf{u}_{k+1}) &= J'(\mathbf{u}_{k+1})(\rho_k \nabla J(\mathbf{u}_k)) + \frac{1}{2} J''\left(\mathbf{u}_k^+\right)(\rho_k \nabla J(\mathbf{u}_k), \rho_k \nabla J(\mathbf{u}_k)) \\
&= J'(\mathbf{u}_{k+1})(\rho_k \nabla J(\mathbf{u}_k)) + \frac{1}{2} J''\left(\mathbf{u}_k^+\right)(\rho_k \nabla J(\mathbf{u}_k), \rho_k \nabla J(\mathbf{u}_k)) \\
&= \langle \nabla J(\mathbf{u}_{k+1}), \rho_k \nabla J(\mathbf{u}_k) \rangle + \frac{\rho_k^2}{2} \left\langle \nabla^2 J\left(\mathbf{u}_k^+\right) \nabla J(\mathbf{u}_k), \nabla J(\mathbf{u}_k) \right\rangle \\
&= 0 + \frac{\rho_k^2}{2} \left\langle \nabla^2 J\left(\mathbf{u}_k^+\right) \nabla J(\mathbf{u}_k), \nabla J(\mathbf{u}_k) \right\rangle
\end{aligned}
$$

Se deduce que:

$$
\begin{aligned}
& \rho_k \|\nabla J(\mathbf{u}_k)\|^2 \leq \frac{\rho_k^2}{2} \left\langle \nabla^2 J\left(\mathbf{u}_k^-\right) \nabla J(\mathbf{u}_k), \nabla J(\mathbf{u}_k) \right\rangle + \frac{\rho_k^2}{2} \left\langle \nabla^2 J\left(\mathbf{u}_k^+\right) \nabla J(\mathbf{u}_k), \nabla J(\mathbf{u}_k) \right\rangle \\
\Rightarrow\ & \rho_k \|\nabla J(\mathbf{u}_k)\|^2 \leq M \rho_k^2 \|\nabla J(\mathbf{u}_k)\|^2, \text{ porque: } \left|\left\langle \nabla^2 J(\mathbf{v}) \mathbf{w}, \mathbf{w} \right\rangle\right| \leq M \|\mathbf{w}\|^2, \forall \mathbf{v} \in U, \mathbf{w} \in \mathbb{R}^n \\
\Rightarrow\ & \frac{1}{M\rho_k} \leq 1
\end{aligned}
$$

de allí $M\rho_k \leq 1$ y con mayor motivo $\tau(\mathbf{u}_k) \leq \frac{1}{M}$ entonces se tiene:

$$J(\mathbf{u}_{k+1}) = \inf_{0\leq\rho\leq\tau(\mathbf{u}_k)} \varphi_k(\rho) \leq \varphi_k\left(\frac{1}{M}\right) = J\left(\mathbf{u}_k - \frac{1}{M}\nabla J(\mathbf{u}_k)\right)$$

de ahí, tras una nueva aplicación de la fórmula de Taylor-MacLaurin, la existencia de un punto $\mathbf{u}_k^* \in \left[\mathbf{u}_k, \mathbf{u}_k - \frac{1}{M}\nabla J(\mathbf{u}_k)\right]$ tal que:

$$\begin{aligned}
J(\mathbf{u}_{k+1}) - J(\mathbf{u}_k) &= J'(\mathbf{u}_k)\left(-\frac{1}{M}\nabla J(\mathbf{u}_k)\right) + \frac{1}{2}J''(\mathbf{u}_k^*)\left(-\frac{1}{M}\nabla J(\mathbf{u}_k), -\frac{1}{M}\nabla J(\mathbf{u}_k)\right) \\
&= \left\langle \nabla J(\mathbf{u}_k), -\frac{1}{M}\nabla J(\mathbf{u}_k)\right\rangle + \frac{1}{2M^2}\left\langle \nabla^2 J(\mathbf{u}_k^*)\nabla J(\mathbf{u}_k), \nabla J(\mathbf{u}_k)\right\rangle \\
&= -\frac{1}{M}\|\nabla J(\mathbf{u}_k)\|^2 + \frac{1}{2M^2}\left\langle \nabla^2 J(\mathbf{u}_k^*)\nabla J(\mathbf{u}_k), \nabla J(\mathbf{u}_k)\right\rangle \\
&\leq -\frac{1}{M}\|\nabla J(\mathbf{u}_k)\|^2 + \frac{1}{2M^2}M\|\nabla J(\mathbf{u}_k)\|^2
\end{aligned}$$

entonces

$$J(\mathbf{u}_{k+1}) - J(\mathbf{u}_k) \leq -\frac{1}{2M}\|\nabla J(\mathbf{u}_k)\|^2$$

(*iv*) Siendo la sucesión $\{J(\mathbf{u}_k)\}_{k\geq 0}^{\infty}$ decreciente, la reducción establecida en (*iii*) implica

$$J(\mathbf{u}_k) - J(\mathbf{u}_{k+p}) \geq \frac{\|\nabla J(\mathbf{u}_k)\|^2}{2M}, \forall p \in \mathbb{N}$$

Para cualquier sucesión extraída $(\mathbf{u}'_k)$ de la sucesión $(\mathbf{u}_k)$ que converge a un punto $\mathbf{u} \in \mathbb{R}^n$, se está pasando por tanto en el límite $\nabla J(\mathbf{u}) = 0$. La existencia de una sucesión extraída convergente en U resulta de la hipótesis de compacidad de U en $\mathbb{R}^n$

(*v*) Sea $\mathbf{v}$ un punto en U distinto de $\mathbf{u}$; la hipótesis adicional de esta pregunta implica en particular que $\nabla J(\mathbf{u}) \neq 0$. Desde el punto $\mathbf{v}_0 = \mathbf{v}$, el algoritmo definido en (*ii*) permite ya sea encontrar un punto $\mathbf{v}_N \in U$ tal que $\nabla J(\mathbf{v}_N) = 0$ o construir una sucesión $(\mathbf{v}_k)_{k\leq 0}$. En el primer caso, se tiene $\mathbf{v}_N = \mathbf{u}$; en el segundo caso, se puede extraer una subsucesión $(\mathbf{v}'_k)$ que converge en U hasta un punto donde ∇J desaparece, por lo tanto converge necesariamente a $\mathbf{u}$. Se tiene en todos los casos

$$J(\mathbf{v}) = J(\mathbf{v}_0) < J(\mathbf{v}_1) \leq J(\mathbf{u})$$

lo que establece que $\mathbf{u}$ es un mínimo estricto del funcional J.

Finalmente, si la sucesión $(\mathbf{u}_k)$ no converge a $\mathbf{u}$, existiría un número $\varepsilon > 0$ y una sucesión extraída $(\mathbf{u}_{k^*})$ tal que $\|\mathbf{u}_k^* - \mathbf{u}\| \geq \varepsilon$ para todo k^*. Como en (*iv*), entonces se puede extraer de la sucesión $(\mathbf{u}_{k^*})$ una sucesión extraída $(\mathbf{u}_{k^{**}})$ convergente en U hacia un punto donde ∇J se anula, por lo tanto converge hacia $\mathbf{u}$. Se deduce por absurdo que toda la sucesión $(\mathbf{u}_k)$ converge a $\mathbf{u}$.

$\boxtimes$

Ejercicio 7.2: El objetivo del problema es estudiar dos *métodos de minimización para funciones de una variable*, que no utilizan la evaluación de la primera derivada (la derivación "numérica" debe evitarse siempre que sea posible). Para lo que sigue se tiene: Sea f una función con valores reales, dos veces continuamente derivable en un intervalo compacto $[a, b] \subset \mathbb{R}$,

tal que $f''(\rho) > 0$ para todo $\rho \in [a,b]$; se nota de pasada que esta última hipótesis se satisface seguramente si la función f es de la forma:

$$f(\rho) = J(\mathbf{w} + \rho\mathbf{d}), \mathbf{d} \neq 0$$

se supone que el funcional J es elíptico y dos veces continuamente derivable. Finalmente se asume la existencia de un punto $c \in]a,b[$ tal que $f'(c) = 0$, que se quiere "ubicar" lo mejor posible (tal punto c es único ya que la función f es estrictamente convexa.

(*i*) Sean x_1 y x_2 dos números tales que $a \leq x_1 < x_2 \leq b$. Demostrar las implicaciones:

$$\begin{aligned} f(x_1) \geq f(x_2) &\Rightarrow x_1 < c < b \\ f(x_2) \geq f(x_1) &\Rightarrow a < c < x_2 \end{aligned}$$

(*ii*) Dado un número arbitrario $\varepsilon > 0$, mostrar que se puede ubicar el punto c en un intervalo de longitud $\leq \left(\frac{b-a}{2} + \varepsilon\right)$ con solo dos evaluaciones de f, por ejemplo $f\left(\frac{a+b}{2}\right)$ y $f\left(\frac{a+b}{2} + \varepsilon\right)$

(*iii*) Demuestre que se puede ubicar el punto c en un intervalo de longitud $\leq \left(\frac{b-a}{3} + \varepsilon\right)$ con solo tres evaluaciones de f. Para eso se comienza comparando los valores $f\left(\frac{2a+b}{3}\right)$ y $f\left(\frac{a+2b}{3}\right)$, luego se vuelve a la cuestión (*ii*)

(*iv*) Los *números de Fibonacci* definidos por recurrencia por la fórmula:

$$\begin{cases} u_0 = 0 \\ u_1 = 1 \\ u_n = u_{n-1} + u_{n-2} \quad, n \geq 2 \end{cases}$$

Dado un entero $n \geq 2$, se establecen:

$$\begin{aligned} x_1 &= a\frac{u_n}{u_{n+1}} + b\frac{u_{n-1}}{u_{n+1}} \\ x_2 &= a\frac{u_{n-1}}{u_{n+1}} + b\frac{u_n}{u_{n+1}} \end{aligned}$$

Verificando las relaciones:

$$\begin{aligned} x_1\frac{u_{n-1}}{u_n} + b\frac{u_{n-2}}{u_n} &= x_2 \\ a\frac{u_{n-2}}{u_n} + x_2\frac{u_{n-1}}{u_n} &= x_1 \end{aligned}$$

Deducir que se puede ubicar el punto c en un intervalo de longitud $\leq \left(\frac{b-a}{u_{n+1}} + \varepsilon\right) = \Delta_n, n \geq 2$, con solo n evaluaciones de f.

(*v*) Se define el *número de oro*: $\phi = \frac{1+\sqrt{5}}{2}$ (se nota que satisface la ecuación $\phi^2 = \phi + 1$), luego los números

$$\begin{aligned} x_1 &= a(\phi - 1) + b(2 - \phi) \\ x_2 &= a(2 - \phi) + b(\phi - 1) \end{aligned}$$

Calcule $(b - x_1)$ y $(x_2 - a)$; deducir que se puede ubicar el punto c en un intervalo de longitud $(b-a)(\phi - 1)$ usando evaluaciones $f(x_1)$ y $f(x_2)$

(*vi*) Verificar las relaciones:

$$\begin{aligned} x_1 &= a\,(2-\phi) + x_2\,(\phi - 1) \\ x_2 &= x_1\,(\phi - 1) + b\,(2-\phi) \end{aligned}$$

deducir que se puede ubicar el punto c en un intervalo de longitud $(b-a)(\phi-1)^{n-1} = \delta_n, n \geq 2$, con solo n evaluaciones de f.

(*vii*) En la parte (*iv*): calcular u_n en función de n

(*viii*) Demuestre que si n es lo suficientemente grande y ε lo suficientemente pequeño, la relación $\frac{\delta_n}{\Delta_n}$ es un poco diferente de 1,17.

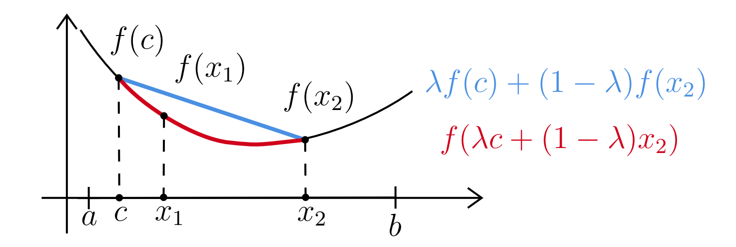
(*ix*) Aunque el primer método es teóricamente mejor, es sin embargo el segundo el que se prefiere utilizar en la práctica. Explicar por qué.

∘ *Solución*

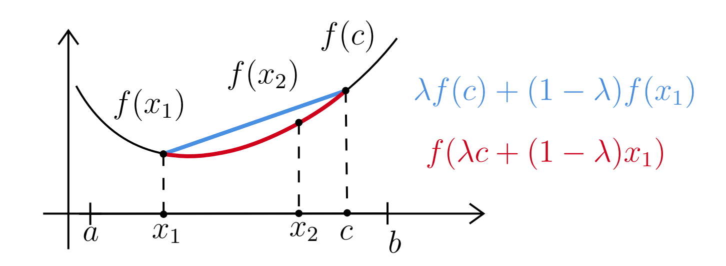
(*i*) Se tiene que $f(x_1) \geq f(x_2)$ y hay que probar que $x_1 < c < b$, se sabe que $c \in]a,b[$, por lo que se suponen $a < c < x_1 < x_2 < b$; entonces $x_1 = \lambda c + (1-\lambda)x_2$, para $0 < \lambda < 1$ y la convexidad estricta de f implica que:

$$\begin{aligned} f(x_1) = \quad f(\lambda c + (1-\lambda)x_2) &< \lambda f(c) + (1-\lambda) f(x_2) \\ \Rightarrow f(x_1) - f(x_2) &< \lambda\,[f(c) - f(x_2)] \\ \Rightarrow f(x_1) - f(x_2) &< 0 \\ \Rightarrow f(x_1) &< f(x_2) \end{aligned}$$

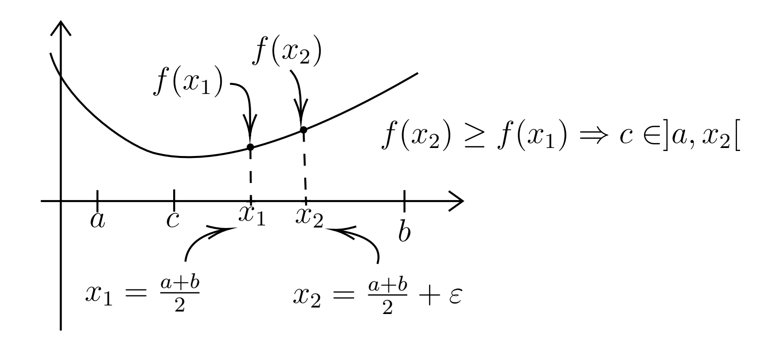
, esto es porque c el mínimo de f pero se tiene una contradicción con la hipótesis $f(x_1) \geq f(x_2)$ por lo cual $x_1 < c < b$ como de quería demostrar; gráficamente se nota que c no puede ser mínimo de f ni f estrictamente convexa al mismo tiempo bajo las condiciones consideradas:



De manera similar se puede demostrar que si $f(x_2) \geq f(x_1) \Rightarrow a < c < x_2$, se asume que $a < x_1 < x_2 < c < b$, gráficamente la contradicción es evidente:

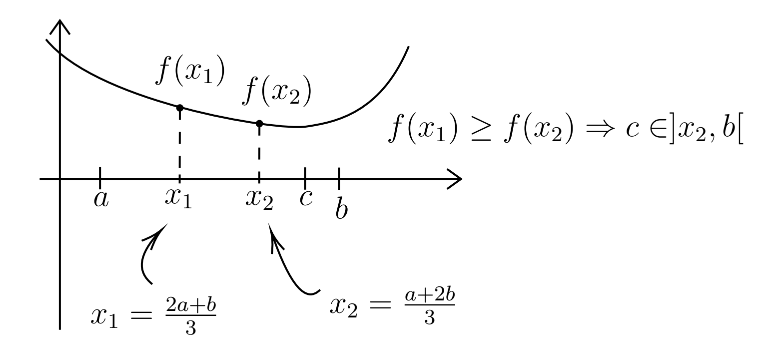


(*ii*) Dado un $\varepsilon > 0$, hay que ver que se puede hubicar el punto c en un intervalo de longitud $\leq \frac{b-a}{2} + \varepsilon$. Sean $x_1 = \frac{a+b}{2}$ y $x_2 = \frac{a+b}{2} + \varepsilon$; si $f(x_2) \geq f(x_1)$, entonces por (*i*) el punto c pertenece al intervalo $]a, x_2[$, que es de longitud $\frac{b-a}{2} + \varepsilon$, gráficamente:

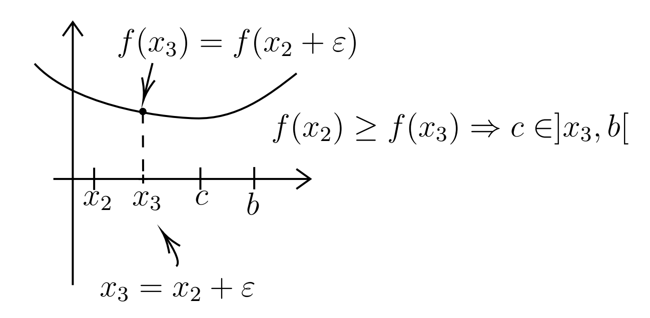


Notar que si $f(x_1) \geq f(x_2)$ se tiene por (*i*) que $c \in]x_1, b[$, que es de longitud menor que $\frac{b-a}{2} + \varepsilon$.

(*iii*) Sean $x_1 = \frac{2a+b}{3}$ y $x_2 = \frac{a+2b}{3}$; si $f(x_1) \geq f(x_2)$, el punto c pertenece al intervalo $[x_1, b[$ que es de longitud $\frac{2}{3}(b-a)$, gráficamente:



, y como $x_2 = \frac{b+x_1}{2}$ basta evaluar en un punto $x_3 = x_2 + \varepsilon$ para poder ubicar el punto c, si $f(x_2) \leq f(x_3) \Rightarrow c \in]x_3, b[$ y si $f(x_3) \leq f(x_2) \Rightarrow c \in]x_1, x_3 + \varepsilon[$, en todo caso c está en un intervalo de longitud $\leq \left(\frac{b-a}{3} + \varepsilon\right)$, gráficamente:



Si $f(x_1) \leq f(x_2)$, basta con evaluar $x_3 = x_1 + \varepsilon$ para ubicar el punto c en un intervalo de la misma longitud $\frac{b-a}{3} + \varepsilon$

(*iv*) Se analiza en *método de búsqueda de fibonacci*. Sean u_n la sucesión *fibonacci* y $[a, b]$ un intervalo, luego para $n \geq 2$ se definen:

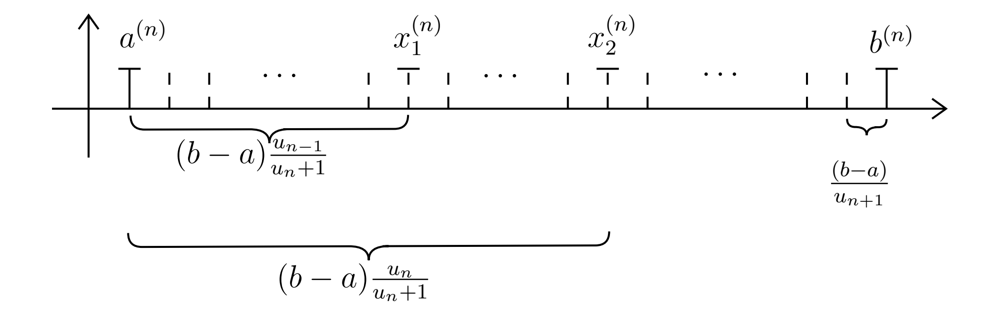
$$\begin{aligned} x_1 &= a\frac{u_n}{u_{n+1}} + b\frac{u_{n-1}}{u_{n+1}} \\ x_2 &= a\frac{u_{n-1}}{u_{n+1}} + b\frac{u_n}{u_{n+1}} \end{aligned}$$

Hay que mostrar que se puede ubicar el punto c en un intervalo de longitud $\leq \left(\frac{b-a}{u_{n+1}} + \varepsilon\right) = \Delta_n, n \geq 2$, con solo n evaluaciones de f.

Con $n = 2$ se tiene el mismo punto, es decir $x_1 = x_2$, se evalúan entonces sólo dos puntos (el otro sería $x_1 + \varepsilon$) y en intervalo tiene $\frac{(b-a)}{2}$ de longitud. Desde el rango $n = 3$ todos los números Fibonacci son estrictamente positivos y $a < x_1 < x_2 < b$. Se plantean:

$$
\begin{aligned}
a^{(n)} &= a \\
b^{(n)} &= b \\
x_1^{(n)} &= a^{(n)} \frac{u_n}{u_{n+1}} + b^{(n)} \frac{u_{n-1}}{u_{n+1}} \\
x_2^{(n)} &= a^{(n)} \frac{u_{n-1}}{u_{n+1}} + b^{(n)} \frac{u_n}{u_{n+1}}
\end{aligned}
$$

, gráficamente:



A partir de las dos evaluaciones: $f\left(x_1^{(n)}\right)$ y $f\left(x_2^{(n)}\right)$, se puede decidir si el punto c pertenece al intervalo $\left[x_1^{(n)}, b^{(n)}\right[$ o al intervalo $\left]a^{(n)}, x_2^{(n)}\right]$, la longitud del segundo intervalo es:

$$
\begin{aligned}
x_2^{(n)} - a^{(n)} &= a^{(n)} \frac{u_{n-1}}{u_{n+1}} + b^{(n)} \frac{u_n}{u_{n+1}} - a^{(n)} \\
&= \frac{au_{n-1} + bu_n - au_{n+1}}{u_{n+1}} \\
&= \frac{au_{n-1} + bu_n - a(u_n + u_{n-1})}{u_{n+1}} \\
&= (b-a)\frac{u_n}{u_{n+1}}
\end{aligned}
$$

, análogamente para el primer intervalo se nota que ambos tienen la misma longitud $(b-a)\frac{u_n}{u_{n+1}}$.

Sean $a^{(n-1)} = x_1^{(n)}, b^{(n)} = b^{(n-1)}$ en el primer caso y $a^{(n-1)} = a^{(n)}, b^{(n-1)} = x_2^{(n)}$ en el segundo. Se puede proceder de la misma forma en el nuevo intervalo $\left[a^{(n-1)}, b^{(n-1)}\right]$, y se introducen los dos puntos:

$$
\begin{aligned}
x_1^{(n-1)} &= a^{(n-1)} \frac{u_{n-1}}{u_n} + b^{(n-1)} \frac{u_{n-2}}{u_n} \\
x_2^{(n-1)} &= a^{(n-1)} \frac{u_{n-2}}{u_n} + b^{(n-1)} \frac{u_{n-1}}{u_n}
\end{aligned}
$$

, para el segundo caso $\left(a^{(n-1)} = a^{(n)} = a, b^{(n-1)} = x_2^{(n)}\right)$, se tiene:

$$\begin{aligned}
x_2^{(n-1)} &= a^{(n-1)}\tfrac{u_{n-2}}{u_n} + b^{(n-1)}\tfrac{u_{n-1}}{u_n} \\
&= a\tfrac{u_{n-2}}{u_n} + \left(a\tfrac{u_{n-1}}{u_{n+1}} + b\tfrac{u_n}{u_{n+1}}\right)\tfrac{u_{n-1}}{u_n} \\
&= a\tfrac{u_{n-2}}{u_n} + a\tfrac{u_{n-1}}{u_{n+1}}\tfrac{u_{n-1}}{u_n} + b\tfrac{u_n}{u_{n+1}}\tfrac{u_{n-1}}{u_n} \\
&= a\tfrac{(u_{n-2}u_{n+1}+u_{n-1}u_{n-1})}{u_{n+1}u_n} + b\tfrac{u_{n-1}}{u_{n+1}} \\
&= a\tfrac{(u_{n-2}(u_n+u_{n-1})+u_{n-1}u_{n-1})}{u_{n+1}u_n} + b\tfrac{u_{n-1}}{u_{n+1}} \\
&= a\tfrac{(u_{n-2}u_n+u_{n-2}u_{n-1}+u_{n-1}u_{n-1})}{u_{n+1}u_n} + b\tfrac{u_{n-1}}{u_{n+1}} \\
&= a\tfrac{((u_n-u_{n-1})u_n+(u_n-u_{n-1})u_{n-1}+u_{n-1}u_{n-1})}{u_{n+1}u_n} + b\tfrac{u_{n-1}}{u_{n+1}} \\
&= a\tfrac{(u_nu_n-u_{n-1}u_n+u_nu_{n-1}-u_{n-1}u_{n-1}+u_{n-1}u_{n-1})}{u_{n+1}u_n} + b\tfrac{u_{n-1}}{u_{n+1}} \\
&= a\tfrac{u_nu_n}{u_{n+1}u_n} + b\tfrac{u_{n-1}}{u_{n+1}} = a\tfrac{u_n}{u_{n+1}} + b\tfrac{u_{n-1}}{u_{n+1}} = x_1^{(n)}
\end{aligned}$$

; análogamente para el primer caso se nota que el punto $x_1^{(n-1)}$ coincide con el punto $x_2^{(n)}$. Suponiendo que $f\left(x_2^{(n-1)}\right) \leq f\left(x_1^{(n-1)}\right)$ se tiene que $a^{(n-2)} = x_1^{(n-1)}$ y $b^{(n-2)} = b^{(n-1)}$, se calcula la longitud de este intervalo:

$$\begin{aligned}
b^{(n-2)} - a^{(n-2)} &= b^{(n-1)} - x_1^{(n-1)} \\
&= a\tfrac{u_{n-1}}{u_{n+1}} + b\tfrac{u_n}{u_{n+1}} - a\tfrac{u_{n-1}}{u_n} - a\tfrac{u_{n-1}}{u_{n+1}}\tfrac{u_{n-2}}{u_n} - b\tfrac{u_n}{u_{n+1}}\tfrac{u_{n-2}}{u_n} \\
&= a\left(\tfrac{u_{n-1}}{u_{n+1}} - \tfrac{u_{n-1}}{u_n} - \tfrac{u_{n-1}}{u_{n+1}}\tfrac{u_{n-2}}{u_n}\right) + b\left(\tfrac{u_n}{u_{n+1}} - \tfrac{u_{n-2}}{u_{n+1}}\right) \\
&= a\left(\tfrac{u_{n-1}u_n-u_{n-1}u_{n+1}-u_{n-1}u_{n-2}}{u_{n+1}u_n}\right) + b\left(\tfrac{u_n-u_{n-2}}{u_{n+1}}\right) \\
&= a\tfrac{(u_{n-1}(u_{n-2}+u_{n-1})-u_{n-1}(u_n+u_{n-1})-u_{n-1}u_{n-2})}{u_{n+1}u_n} + b\tfrac{u_{n-1}}{u_{n+1}} \\
&= a\tfrac{(u_{n-2}u_{n-1}+u_{n-1}u_{n-1}-u_nu_{n-1}-u_{n-1}u_{n-1}-u_{n-1}u_{n-2})}{u_{n+1}u_n} + b\tfrac{u_{n-1}}{u_{n+1}} \\
&= a\tfrac{(-u_nu_{n-1})}{u_{n+1}u_n} + b\tfrac{u_{n-1}}{u_{n+1}} \\
&= (b-a)\tfrac{u_{n-1}}{u_{n+1}}
\end{aligned}$$

, análogamente para el caso en que $f\left(x_2^{(n-1)}\right) \geq f\left(x_1^{(n-1)}\right)$, se tiene que tres evaluaciones son suficientes para reducir el intervalo $\left[a^{(n-2)}, b^{(n-2)}\right]$, de longitud $(b-a)\frac{u_{n-1}}{u_{n+1}}$: Recursivamente basta con $n-1$ evaluaciones para reducir la longitud del intervalo en uno de dos puntos:

$$x_1^{(3)} = \frac{2a^{(3)}+b^{(3)}}{3} \quad \text{o} \quad x_2^{(3)} = \frac{a^{(3)}+2b^{(3)}}{3}$$

Se tiene, por tanto, el caso de la pregunta (iii), y una evaluacion adicional es suficiente, es decir, n en total, para ubicar el punto c en un intervalo de longitud $\leq \left(\frac{b-a}{u_{n+1}} + \varepsilon\right) = \Delta_n$

(v) Se analiza el *método de la sección áurea*. Se tiene que:

$$\begin{aligned}
x_1 &= a(\phi-1) + b(2-\phi) \\
x_2 &= a(2-\phi) + b(\phi-1)
\end{aligned}$$

; donde ϕ es el *número de oro* (*razón áurea*), luego hay que calcular: $(b-x_1)$ y (x_2-a) y deducir que se puede ubicar el punto c en un intervalo de longitud $(b-a)(\phi-1)$ usando evaluaciones $f(x_1)$ y $f(x_2)$. Se nota $a < x_1 < x_2 < b$, luego:

$$\begin{aligned}
b - x_1 &= b - a\phi + a - 2b + b\phi \\
&= \phi(b-a) - (b-a) \\
&= (b-a)(\phi-1)
\end{aligned}
\qquad
\begin{aligned}
x_2 - a &= 2a - a\phi + b\phi - b - a \\
&= a - b - \phi(a-b) \\
&= (a-b)(1-\phi) \\
&= (b-a)(\phi-1)
\end{aligned}$$

Por lo cual: $b - x_1 = x_2 - a = (b-a)(\phi - 1)$

(vi) Con los elementos de la pregunta (v), hay que deducir que se puede ubicar el punto c en un intervalo de longitud $(b-a)(\phi-1)^{n-1} = \delta_n, n \geq 2$, con solo n evaluaciones de f. Procediendo como en la pregunta (iv): Si en el paso i el punto está ubicado en el intervalo $\left[a^{(i)}, b^{(i)}\right]$, entonces en el paso *i+1* se ubicará ya sea en el intervalo $\left[a^{(i+1)} = x_1^{(i)}, b^{(i+1)} = b^{(i)}\right]$ y en este caso $x_1^{(i+1)} = x_2^{(i)}$; ya sea en el intervalo $\left[a^{(i+1)} = a^{(i)}, b^{(i+1)} = x_2^{(i)}\right]$ y en este caso $x_2^{(i+1)} = x_1^{(i)}$; cada uno de estos intervalos tiene una longitud $\left(b^{(i)} - a^{(i)}\right)(\phi - 1)$. Con n evaluaciones de f, se puede ubicar el punto c en un intervalo de longitud $\leq \left(b^{(n-1)} - a^{(n-1)}\right)(\phi-1) = \left(b^{(1)} - a^{(1)}\right)(\phi-1)^{n-1} = (b-a)(\phi-1)^{n-1} = \delta_n$

(vii) Se observa que u_n es el término general de una recurrencia lineal; por lo tanto $u_n = \alpha r^n + \beta q^n$, donde r y q son raíces de $r^2 - r - 1$, por lo tanto:

$$\begin{aligned} r &= \tfrac{-(-1)\pm\sqrt{(-1)^2-4(1)(-1)}}{2(1)} \\ &= \tfrac{1\pm\sqrt{5}}{2} \end{aligned}$$

$$\begin{aligned} \Rightarrow\ r &= \phi \\ \Rightarrow\ q &= \tfrac{1-\sqrt{5}}{2} = \tfrac{1}{2} - \tfrac{\sqrt{5}}{2} = 1 - \tfrac{1}{2} - \tfrac{\sqrt{5}}{2} = 1 - \phi \end{aligned}$$

Las constantes α y β están determinadas por las condiciones iniciales $u_0 = 0, u_1 = 1$:

$$\begin{aligned} u_n &= \alpha r^n + \beta q^n \\ &= \alpha\phi^n + \beta(1-\phi)^n \end{aligned}$$

$$\begin{aligned} u_0 &= \alpha + \beta = 0 \\ \Rightarrow\ \alpha &= -\beta \end{aligned}$$

$$\begin{aligned} u_1 &= \alpha\phi - \alpha(1-\phi) = 1 \\ \Rightarrow\ \alpha &= \tfrac{1}{(\phi-1+\phi)} \\ &= \tfrac{1}{2\phi-1} \\ &= \tfrac{1}{1+\sqrt{5}-1} \\ &= \tfrac{1}{\sqrt{5}} \\ \Rightarrow\ \beta &= -\tfrac{1}{\sqrt{5}} \end{aligned}$$

Finalmente, se obtiene:

$$u_n = \frac{1}{\sqrt{5}}\left[\phi^n - (1-\phi)^n\right]$$

$(viii)$ Para ε lo suficientemente pequeño:

$$\delta_n = (b-a)(\phi-1)^{n-1} \qquad \Delta_n = \tfrac{b-a}{u_{n+1}}$$

$$\begin{aligned} \Rightarrow\ \tfrac{\delta_n}{\Delta_n} &= u_{n+1}(\phi-1)^{n-1} \\ &= \tfrac{1}{\sqrt{5}}\left(\phi^{n+1} - (1-\phi)^{n+1}\right)(\phi-1)^{n-1} \\ &\approx \tfrac{1}{\sqrt{5}}\phi^{n+1}(\phi-1)^{n-1}\ ,\ 0 < 1-\phi < 1 \\ &= \tfrac{1}{\sqrt{5}}\phi^2(\phi(\phi-1))^{n-1} \\ &= \tfrac{1}{\sqrt{5}}\phi^2(\phi^2-\phi)^{n-1} \\ &= \tfrac{\phi^2}{\sqrt{5}}\ ,\ \phi^2 - \phi - 1 = 0 \\ &\approx 1{,}17 \end{aligned}$$

(ix) Aquí hay que comparar los dos métodos definidos. Con el primer método (*búsqueda de fibonacci*), para construir los dos primeros puntos x_1 y x_2, es necesario conocer de antemano el número de evaluaciones de f que se aceptan realizar, es decir la precisión con la que se quiere ubicar el punto c. Por lo tanto, se prefiere en la práctica el segundo método (*sección áurea*).

⊠

Ejercicio 7.3: Describir la aplicación del método de gradiente con paso óptimo al funcional:

$$J : \mathbf{v} = (v_1, v_2) \in \mathbb{R}^2 \longrightarrow J(\mathbf{v}) = v_1^4 - 4v_1^3 + 6\left(v_1^2 + v_2^2\right) - 4(v_1 + v_2)$$

explicando en particular la ecuación escalar a resolver en cada iteración. ¿Los resultados del material (existencia de un mínimo, convergencia del método, etc...) se aplican a este ejemplo?

∘ *Solución*

Para la primera parte, se denota con f el funcional, para aplicar el método de gradiente con paso óptimo se calcula el gradiente del funcional:

$$\nabla f(v_1, v_2) = \left(4v_1^3 - 12v_1^2 + 12v_1 - 4, 12v_2 - 4\right)$$

luego se plantea el algoritmo correspondiente:

$$\begin{cases} \mathbf{v}^0 & \in & \mathbb{R}^2 \\ \mathbf{v}^k & = & \mathbf{v}^{k-1} - \varrho_{k-1}\nabla f\left(\mathbf{v}^{k-1}\right), k \geq 1 \\ & = & \left(v_2^{k-1}, v_2^{k-1}\right) - \varrho_{k-1}\left(4\left(v_1^{k-1}\right)^3 - 12\left(v_1^{k-1}\right)^2 + 12v_1^{k-1} - 4, 12v_2^{k-1} - 4\right) \end{cases}$$

; donde cada ϱ_{k-1} es tal que para una función $h : \mathbb{R} \longrightarrow \mathbb{R}$:

$$\begin{cases} h(\varrho_{k-1}) & = & f\left(\mathbf{v}^{k-1} - \varrho_{k-1}\nabla f\left(\mathbf{v}^{k-1}\right)\right) \\ h'(\varrho_{k-1}) & = & 0 \end{cases}$$

, y además se toma el ϱ_{k-1} que genere el menor valor en h, para ello se pueden aplicar otros métodos de aproximación dado que la expresión de la función h puede ser muy compleja.
Para la segunda parte, notar que dada la forma del funcional es claro que no es cuadrática pero se revisa si es elíptica, es decir cumple que dado $J : V \to \mathbb{R}$ definida en un espacio de Hilbert V, J se llama *elíptico* si es continuamente derivable en V y si existe una constante $\alpha > 0$, tal que

$$\alpha > 0 \text{ y } \langle \nabla J(\mathbf{v}) - \nabla J(\mathbf{u}), \mathbf{v} - \mathbf{u}\rangle \geq \alpha \left\|\mathbf{v} - \mathbf{u}\right\|^2 \text{ para todo } \mathbf{u}, \mathbf{v} \in V$$

Notar que para f, se tiene que $\mathbb{R}^2$ es un espacio de Hilbert, y además derivable, ahora se busca generar la desigualdad: dados $\mathbf{v} = (x, y)$ y $\mathbf{u} = (a, b)$:

$$\begin{aligned} \langle \nabla J(\mathbf{v}) - \nabla J(\mathbf{u}), \mathbf{v} - \mathbf{u}\rangle &= (4x^3 - 12x^2 + 12x - 4a^3 + 12a^2 - 12a)(x - a) \\ &\quad +(12y - 12b)(y - b) \\ &= 4x^4 - 12x^3 - 4a^3x + 12a^2x - 4ax^3 + 12ax^2 + 4a^4 - 12a^3 \\ &\quad +12(x^2 - 2ax + a^2 + y^2 - 2by + b^2) \\ &= (4(x^2 + ax + a^2 - 3x - 3a) + 12)(x - a)^2 + 12(y - b)^2 \end{aligned}$$

Si existe el $\alpha > 0$ que cumpla la desigualdad de la definición, entonces existe solución de esta inecuación:

$$\begin{aligned} 4\left(x^2 + ax + a^2 - 3x - 3a\right) + 12 &> 0 \\ \Rightarrow x^2 + ax + a^2 - 3x - 3a &> -3 \end{aligned}$$

minimizando la última expresión de forma analítica se tiene que el mínimo está en $(x, a) = (1, 1)$ y es -3, así se tiene que para $\mathbf{v}, \mathbf{u} \in \mathbb{R}^2 - \{(1, y), y \in \mathbb{R}\}$, existe $\alpha > 0$ tal que:

$$\begin{aligned} \langle \nabla J(\mathbf{v}) - \nabla J(\mathbf{u}), \mathbf{v} - \mathbf{u} \rangle &= \left(4\left(x^2 + ax + a^2 - 3x - 3a\right) + 12\right)(x - a)^2 + 12(y - b)^2 \\ &\geq \alpha\left((x - a)^2 + (y - b)^2\right), \alpha > 0 \\ &= \alpha \parallel \mathbf{v} - \mathbf{u} \parallel^2 \end{aligned}$$

por lo que f es elíptico en ese conjunto definido anteriormente. Ahora se considera este resultado:

- ***Teorema 7:***

  (*ii*) Si U es un conjunto no vacío, convexo y cerrado del espacio de Hilbert V, y si J es una función elíptica. Ahora, el problema: Encontar $\mathbf{u}$ tal que

  $$\mathbf{u} \in U \subseteq V \text{ y } J(\mathbf{u}) = \inf_{\mathbf{v} \in U} J(\mathbf{v}) \tag{P}$$

  tiene una solución y sólo una.

Notar que el conjunto $\mathbf{v}, \mathbf{u} \in \mathbb{R}^2 - \{(1, y), y \in \mathbb{R}\}$ parte el plano en dos secciones que cumplen las propiedades de ser no vacías, convexas pero no son cerradas en $\mathbb{R}^2$ por lo cual no se puede aplicar este teorema.

⊠

8. Conclusiones

- El método de relajación es una generalización del método de *Gauss-Stilges* para la solución de sistemas de ecuaciones lineales
- La existencia del gradiente es fundamental para los métodos de gradiente, es aconsejable plantear otras formas para generar los valores del vector gradiente que no impliquen cálculos numéricos, esto para no generar más imprecisiones en las aproximaciones.
- En el método de gradiente conjugado se está generando un sustituto del vector gradiente pero en términos del mismo gradiente, no es una aproximación de este.
- La elección de la magnitud ϱ_k en cada iteración, determina en gran medida la velocidad de convergencia del método
- En el caso de lo métodos con restricciones hay que tomar en cuenta nada más la definición de la función proyección para su ejecución.

9. Bibliografía

- Ciarlet, Philippe G. Introduction à l'analyse numérique matricielle et à l'optimisation